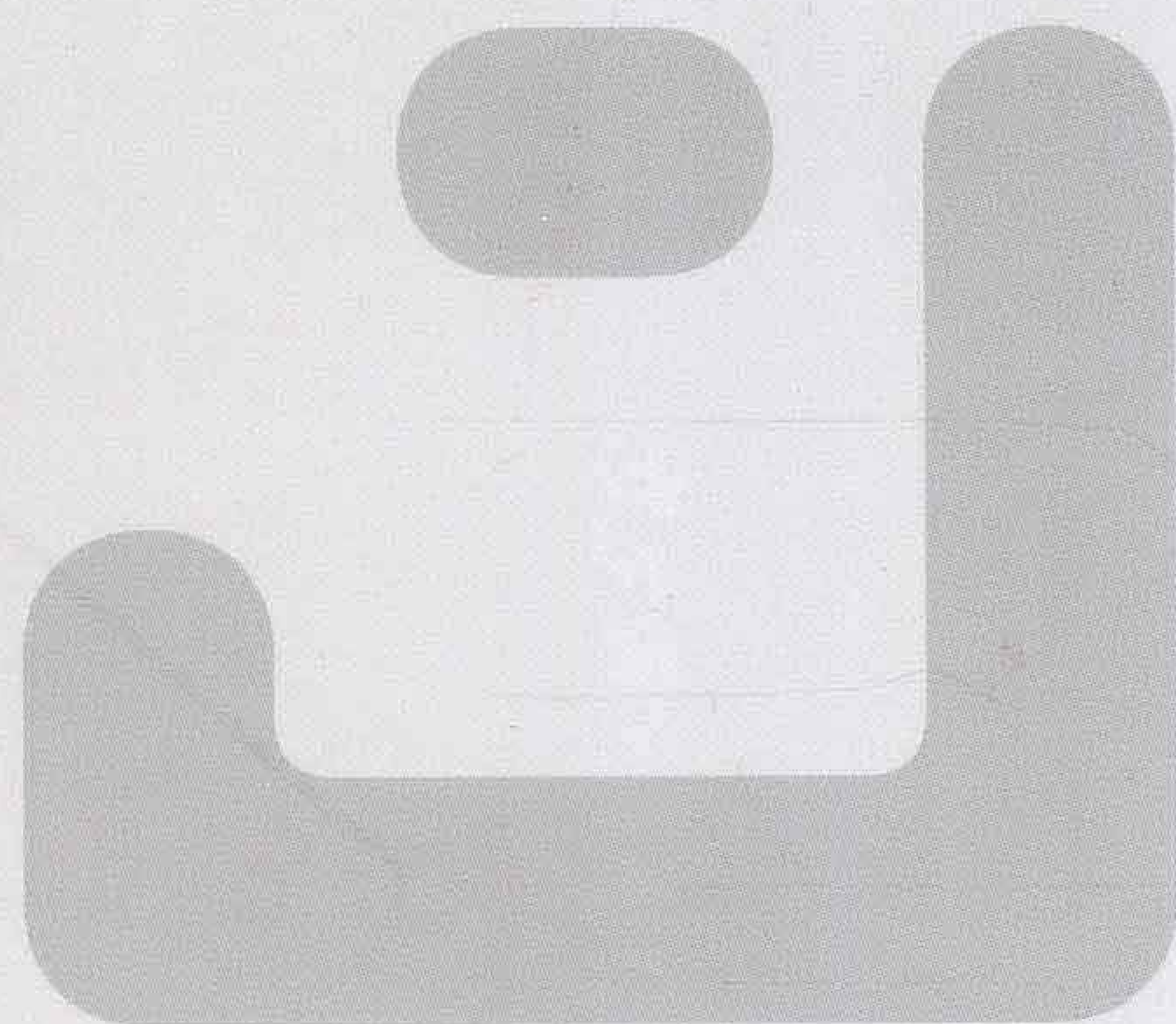
万卷方法



CATEGORICAL
DATA
ANALYSIS

分类数据分析

阿兰·阿格莱斯蒂 著
(Alan Agresti)

齐亚强 译



重庆大学出版社
http://www.cqup.com.cn

FENLEI SHUJU FENXI

由于分类数据分析技术的发展以及分类数据在现实应用中的独特价值，许多统计系或生物统计系都开设了有关分类数据分析的课程。这本书可以用作该类课程的教科书。本书的第1—7章涵盖了该类课程的核心内容。其中，第1—3章介绍分类结果变量的分布以及传统的二维列联表分析方法。第4—7章介绍关于二分和多项分布结果变量的logistic回归以及相应的logit模型。第8章和第9章的内容则是用于分析列联表数据的对数线性模型。随着时间的推移，对数线性模型的重要性似乎有所降低，所以本版在一定程度上缩减了对该模型的讨论，并相应增加了有关logistic回归的内容。

在过去10年间，这一领域的新发展主要集中于对重复测量和其他形式的群组分类数据的分析方法。第10—13章讲述这些方法，其中包括边际模型和具有随机效应的广义线性混合模型。第14—15章介绍本书所使用的最大似然估计的理论基础以及其他可供选择的估计方法。第16章简单回顾了分类数据分析技术的发展历程，并介绍了诸如皮尔逊和费舍尔等著名统计学家的贡献，他们的开创性工作为分类数据分析方法的发展奠定了基础。

本书配套数据及程序可以通过以下途径获取：

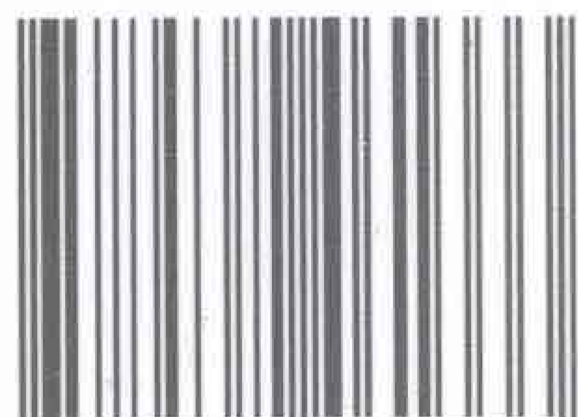
313745784@qq.com（注明书名即可）

http://q.blog.sina.com.cn/fafang/bbs/topic/tid=14632370

参阅及发表相关评论，请登录万卷方法博客圈：

http://q.blog.sina.com.cn/fafang

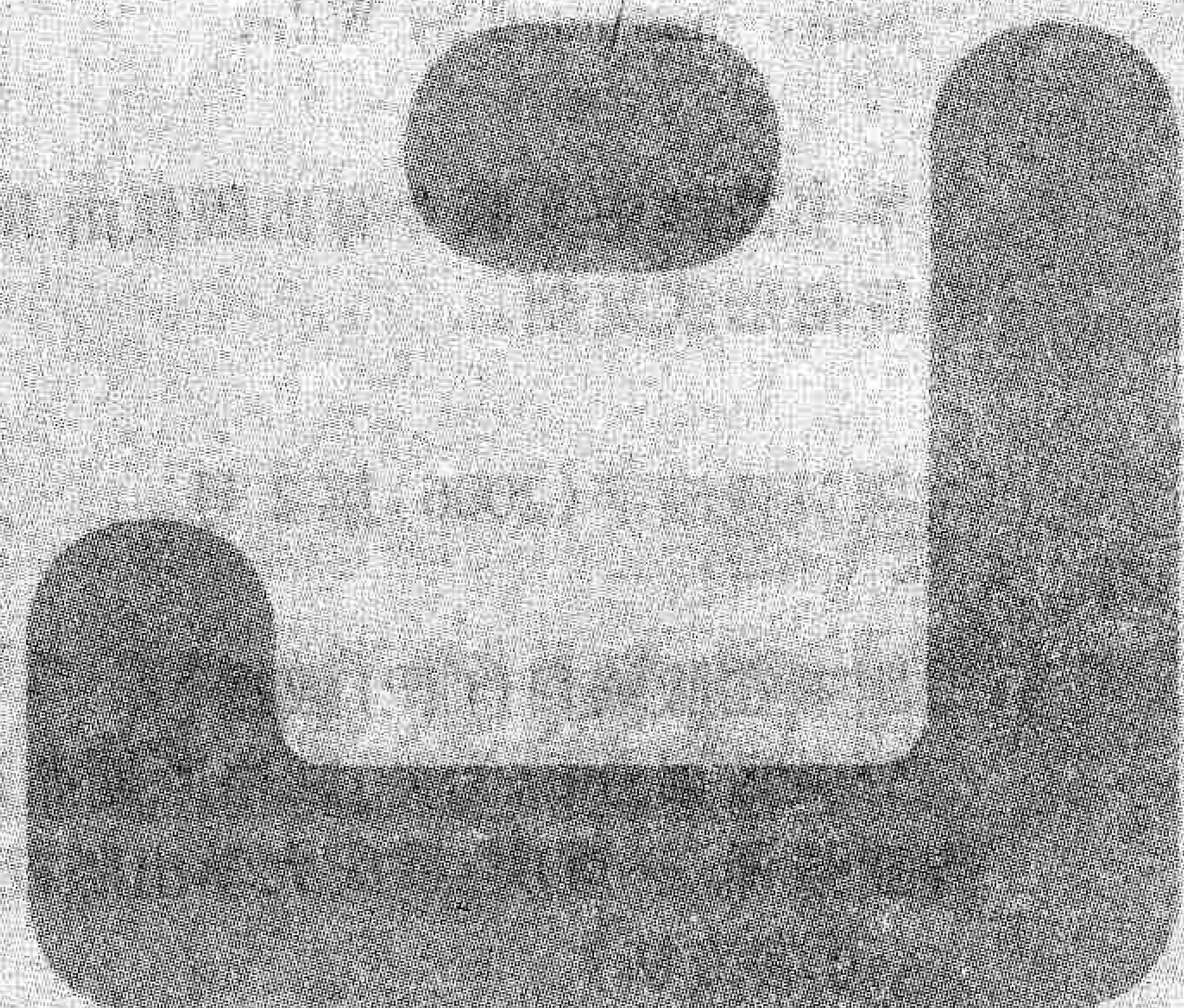
ISBN 978-7-5624-6133-3



定价:82.00元

万卷方法

CATEGORICAL
DATA
ANALYSIS



分类数据分析

阿兰·阿格莱斯蒂 著
(Alan Agresti)

齐亚强 译

重庆大学出版社

Categorical Data Analysis. By:ALAN AGRESTI
ISBN:0471360937

版贸核渝字(2006)第 2 号。

图书在版编目(CIP)数据
分类数据分析/(美)阿格莱斯蒂(Agresti,A.)著;
齐亚强译.—重庆:重庆大学出版社,2012.1
(万卷方法)
书名原文:categorical data analysis
ISBN 978-7-5624-6133-3

Ⅰ.①分… Ⅱ.①阿…②齐… Ⅲ.①统计数据—统计分析 Ⅳ.①O212

中国版本图书馆 CIP 数据核字(2011)第 105669 号

分类数据分析
阿兰·阿格莱斯蒂 著
齐亚强 译
策划编辑:雷少波
责任编辑:文 鹏 罗 杉 版式设计:雷少波
责任校对:邹 忌 责任印制:赵 晟
*
重庆大学出版社出版发行
出版人:邓晓益
社址:重庆市沙坪坝区大学城西路 21 号
邮编:401331
电话:(023)88617183 88617185(中小学)
传真:(023)88617186 88617166
网址:http://www.cqup.com.cn
邮箱:fxk@cqup.com.cn (营销中心)
全国新华书店经销
自贡新华印刷厂印刷
*
开本:787×1092 1/16 印张:32.75 字数:814 千
2012 年 1 月第 1 版 2012 年 1 月第 1 次印刷
印数:1—4 000
ISBN 978-7-5624-6133-3 定价:82.00 元

总序
社会研究方法的现状及其发展趋势

近年来，社会调查技术和社会研究方法都有很大的发展。在调查技术方面，自 20 世纪 70 年代以来，社会变迁多次横断面的跟踪调查研究，几乎成为所有国家和地区了解社会结构转变和社会发展状况的基础性调查。这种调查不仅对社会学的研究有很大促进作用，而且对整个社会科学的研究都产生了重大影响，并且这些调查结果有的已作为政府有关部门决策的重要依据。国际上比较著名的此类调查有：美国芝加哥大学全国民意调查中心（National Opinion Research Center，简称 NORC）的"社会综合调查（General Social Survey ，简称 GSS）"，英国埃塞克斯大学调查中心进行的"全国家庭生活和社会变迁调查"，法国经济和社会调查所进行的"全国经济社会调查"，日本社会学会组织进行的"全国社会分层与社会流动调查（简称 SSM）"。中国台湾"中央"研究院社会学研究所，也每两年进行一次"台湾社会变迁基本调查"。美国的"社会基础调查"，现在已成为年度性的调查项目，它是美国国家基金会目前资助的最大的社会科学研究项目。以上这些调查，除美国的调查外，一般均因经费原因采用纵向的间隔性重复调查法，即每隔一段时间，进行一次全国规模的抽样调查。每次调查除保留社会研究所需的基本项目外，都有不同的主题。在间隔若干时间后，再重复同一主题的调查，这样的研究设计，使社会变迁研究在可以涉及更为广泛的研究领域的同时，具有更好的积累性和可比性。多年来，这些基础性调查获得的资料，滋养着大批的社会科学研究者，有时一项调查就有很多名博士生用来写博士论文，以此取得的研究成就，其可靠性受到社会科学界的广泛认同。例如 1997 年出版，以台湾地区社会变迁基本调查数据为基础的研究报告集《90 年代的台湾社会，社会变迁基本调查研究系列二》收集论文 16 篇，内容涉及社会生活的各个方面，在台湾地区引起了极大的反响。

国内社会科学界在这方面也有了长足的发展。笔者所在的中国社会科学院社会学研究所的社会调查和方法研究室，组织或参与了多项与社会变迁有关的大规模抽样调查，取得了一定的研究成果，并积累了大量有关社会变迁的宝贵数据资料，其中主要有：

1. 城乡家庭变迁系列调查：该课题是由中国社会科学院社会学研究所牵头，联合北京大学和地方社科院的研究人员展开的一项类似多次横断面的城乡家庭变迁调查。这一调查始于 1981 年的"中国五城市婚姻家庭调查"，而后有 1988 年的"中国农村家庭调查"、1991 年的"中国七城市家庭调查"、1998 年的"中国城乡家庭变迁调查"。

2. 有关中国城乡社会变迁的系列调查:这一调查始于1991年的第二批国情调查,然后有1992年的“中国城乡居民生活调查”、1993年的“第三批国情调查”、1995年的“第四批国情调查”和1997年的“中国沿海发达地区社会变迁调查”。上述调查虽然还不是严格意义上的多次横断面的纵贯研究,但研究者已在研究设计中尽量考虑到纵贯研究的基本原则,如调查队伍的稳定、指标的可比性和样本空间的延续性等。

3. 中国城乡社会变迁调查:这一调查始于2000年,为中国社会科学院重大课题。目前已经完成第一期第一次调查和第二次调查,今后将把这一调查发展为连续的、定期进行的社会变迁调查。

在纵向调查技术取得长足进步的同时,20世纪末至今,电话调查也有很大发展。电话调查涉及的范围几乎与个别(面对面)访谈同样全面。电话调查中使用的一系列方法,是在20世纪70年代后期和面对面调查一起发展起来的。在20世纪80年代中期,电话调查开始变得很普遍,并且成为许多场合中各种调查方法的首选。正如某些学者所言,一种在公共和私营部门被人们用来帮助提高决策效率的收集信息的有效方法为人们所普遍认同时,这一现象本身就具有方法论上的意义。不仅如此,电话调查还有很大的实践意义,因为它为研究者提供了更多的控制调查质量的机会。这一机会包括抽样、被调查人的选择、问卷题项的提问、计算机辅助电话访谈(CATI)和数据录入。正因为如此,今天在各种社会调查中,如果没有发现其他重要的足以放弃使用电话调查的原因,电话调查由于其独特的对调查质量进行全面监控的优点,常常成为各种调查方式的首选。由笔者翻译,重庆大学出版社出版的《电话调查方法:抽样、选择和督导》一书,也于2005年面世。

无论是纵向调查抑或电话调查,实际上都是收集研究资料的方法,而应用社会科学的发展,不仅在于调查技术,即收集资料技术的发展,还在于研究方法和分析技术的发展。近年来,无论是定性研究方法,还是定量研究方法都有了长足的发展。

首先,计算机技术的发展可谓突飞猛进,它对当今社会生活的各个方面产生了巨大的影响,在悄悄地改变着社会科学的研究风格和研究方式的同时,也大大提升了社会科学学者的研究能力。这种影响表现在研究过程的各个阶段,从理论建构(概念映射)、问卷设计(专业的问卷设计软件)、调查实施(计算机辅助访谈、计算机辅助电话访问系统、网络在线调查系统)、数据录入(光学标记识别软件)到数据分析(包括文本、声音、图像资料的处理),甚至延伸到写作发表阶段。这样的过程发生在如社会学、经济学、政治学、心理学、教育学中,促进了学科之间的相互借鉴和交叉融合,至少在研究方法上呈现出这种趋势。随着计算机计算能力的大幅度提高,20世纪80年代后期,统计学领域内发生了一场“革命”,主要表现在对定类和定序变量的建模能力的大幅度提高上,以及与分布无关的统计分析模型的发展之上,特别是基于“Resampling”(包括Bootstrap、Jackknife、Monte Carlo模拟等)的建模技术。同时,计算能力的提高还带动了基于神经网络、动态模拟、人工智能、生态进化等新兴的分析和预测模型的发展。这些进展都为定量社会科学研究提供了更多的可供选择的工具。

亚德瑞安·E.拉夫特里(Adrian E. Raftery)依据社会学家所处理的数据类型,将定量社会学在美国的发展划分为三个时代:第一代起始于20世纪40年代,交互表是其主要处理对象,研究重点是关联度和对数线性模型;第二代起始于20世纪60年代,主要处理单层次的调查数据,Lisrel类型的因果模型和事件史分析是其研究重点;第三代起始于20世纪80年代后期,开始处理诸如文本、空间、社会网络等非传统的数据类型,目前尚没有形成成熟的形态。拉夫特里的综述,虽然更强调定量社会学研究对统计学的贡献,但也

大致勾勒出定量社会学在国外的发展脉络。

从分析模型的角度来看,定量分析在以下几个方向有了突破性发展:

1. 缺失值处理:由于社会生活的复杂性,社会调查数据常常出现缺失值,传统的处理方式是忽略这些缺失值,或者用均值替代。但现在则倾向于用多重插值法(multiple implation)或者其他基于模型的方法进行处理。这些技术的发展,不仅会增强我们对数据的处理能力,而且将改变我们设计问卷的方式。基于这些技术,我们在不增加被访者负担的前提下,大大增加了调查问卷的内容:每个被访者只回答问卷的一部分,然后通过对缺失值的处理,获得他们对未回答部分的估值。

2. 非线性关系:线性假定是经典定量分析的一个常见假定,但在实际研究当中,线性假定只能被看作是对社会现实的一个逼近和简化。面对具体的研究数据,如果没有理论上的明确指引(不幸的是,我们常常没有中程理论的指引),我们是无法在线性模型和非线性模型之间作出取舍的。但 MARS 模型的出现,让我们可以从经验数据当中获得最为拟合的变量之间的函数关系,而不必预先作出线性假定。这样,理论思考和数据分析就可以实现一个互动的循环过程,定量分析就不单单是对理论和假设的简单证伪过程,而是理论思维一个重要组成部分。

3. 测量层次:20 世纪六七十年代的统计模型,大多要求数据的测量层次在定距以上,如因素分析,但社会学的调查数据却大多为定类或定序数据。对应分析、Loglinear、Logit、Logistic Regression、潜类分析、Ordinal Regression、Normal Ogive Regression 等统计模型的出现,大大提高了定量社会学处理定类和定序数据的能力。

4. 测量模型:基于文化、社会、心理和认知等方面的考虑,在社会学界仍有人对问卷调查在中国的效度提出质疑。抛弃"本土化"的文化执著,我们更应当关注的是问卷调查的项目反应理论(item response theory),即被访者回答问卷题器时的过程模型。这方面的进展主要表现在两个方面:一是分解测量量表的成分,如 Rasch model、IRT 分析、Mokken 分析等;二是将测量模型与因果模型或其他分析模型结合在一起,明确把测量误差引入到分析当中,充分评估它们对分析结果的影响,如结构方程模型。

5. 潜变量模型:与测量模型相关联的另外一个发展方向是潜变量模型,例如,潜变量分层分析(latent class analysis)、潜变量结构分析(latent structure analysis)、潜变量赋值分析(latent budget analysis)等。"潜变量"这一概念表明,我们可以通过测量"显变量"来测量无法直接观察的理论概念,如权力、声望、地位等。这样,理论和现实之间,通过"潜变量"到"显变量"的映射(测量过程),就有了连接的桥梁。

6. 分析单元的层序性:在定量分析当中,我们常常强调要避免出现"生态谬误",即分析单元的层次和结论或推论的层次不一致。与其相关的方法论争论是"宏观和微观"的问题。随着多层次模型的出现,我们可以同时考察多个层次上的问题,我们可以把个人放在其家庭背景中,再把家庭放在社区的背景下,考察个人层次的变量对社区变量的效应,或者社区层次的变量对个体行为的具体影响。在定量分析模型当中,"宏观和微观"的连接获得了建模技术上的支持。在这个领域当中,还有一个方向也值得关注:分析宏观层次的数据,对微观层次进行推论。

7. 社会网络模型:区分"关系数据"和"属性数据",是把分析重点从个体/群体等社会单元转移到这些社会单元之间关系的第一步,社会网络模型是目前发展较快的一个定量分析领域,其理论根基是结构主义。社会网络分析目前仍然具有较浓厚的"形态学"特征(基于图论的缘故),但却为我们理解社会关系在社会空间上的形态奠定了基础,通过计算机模拟和研究社会网络的历期数据,研究社会结构的"发生学"性质模型也处在萌芽状

态当中。

8. 系统动力学:如果说社会网络模型是在社会空间上拓展定量社会学的研究手段,那么社会过程在时间上和物理空间上的属性,则是事件史模型、事件数模型、历期分析、Cox 回归、时间序列分析、Cohort 分析、状态空间模型等模型的研究对象。在这个领域,计量经济学为定量社会学研究提供了许多有益的范例。

9. 预测模型:上述模型仍然是在分析主义的范式下。有些社会学的应用研究,更强调模型的预测精度,而不是模型的认知价值,例如,社会趋势的预测。由于计算能力的提高,神经网络、基因算法、人工智能、模式识别等数据挖掘技术有了长足发展,已经出现了许多拟合经验数据的预测模型,比较成功的应用出现在计量经济学领域(如对股市的预测)。

10. 计算机模拟:对于社会学应用研究而言,研究的对象具有历史性、规模大、变迁的过程不仅漫长且表现某种渐进性的特点,且因社会隔离/社会伦理原因无法接近或有实验禁忌等,无法直接进行观察和研究,这时计算机模拟就成为一个可供选择的替代方案。计算机模拟主要有两个类型:一是基于计算机网络的模拟:每台微机作为一个代理,整个网络作为“社会”实时演化,如法国的 Swarm 计划;二是基于概念模型的系统,在计算机时间上,按照既定规则运行,较有名的研究是罗马俱乐部的《增长的极限》,常见的软件有 Simul,Arena 等。自然科学家对此方向似乎比社会学家更有兴趣。

定性研究方法一直是社会学研究领域中比较传统的研究方法,在社会学研究的古典时期,它甚至是社会学家手中唯一的研究方法。但随着定量研究方法在社会学研究中的广泛应用,定性研究方法就似乎越来越不受人们的重视。但需要澄清的事实是,在定量分析模型取得飞速发展的同时,在过去的二十多年里,定性研究方法也有了长足的进步。主要表现在以下六个方面:

1. 研究素材日益扩大:除了传统的参与观察、深度访谈、专题小组访谈之外,会话、交谈、电视、广播、文档、日记、叙事、自传(autobiography)等社会过程中自然产生的素材,甚至社会学理论本身(理论的形式化),也开始进入定性分析的视野当中。所有这些资料,不仅可以以文本的格式存储,而且,新型的多媒体介质,如图像、声音和视频,作为原始的分析素材,也日益成为定性分析的新宠。

2. 分析方法更加多样:定性方法的种类在最近的二十多年中,更是有了一个质的飞跃。在比较传统的、源自语言学的方法,如内容分析、话语分析、修辞分析、语意分析、符号学、论据分析等方法之外,社会学家也创造出自己独特的定性分析方法,如施特劳斯(Strauss)等人的扎根理论、海斯(Heise)的事件结构分析、拉津(Ragin)的定性对比分析、Abbott 和 Hrycak 采用最优匹配技术的序列分析、亚贝儿(Abell)的形式叙事分析(formal narrative analysis)、鲍尔(Bauer)等人的语库建设、Attride-Stirling 等人的主题网络分析和神经网络技术应用的定性分析领域。所有这些方法的一个共同特征是,把定性研究向更加系统、更加精确、更加严格、更加形式化的方向推进。

3. 认识论基础更加多元化:现象学、释义学和本土方法论(ethnomethodology)的认识论,一直是定性分析的大本营,但近年来,实证主义也开始逐渐为定性分析所接纳,解释和阐释之间,由激烈的对立关系,逐渐演变为相互融洽的关系。

4. 研究过程更加客观规范:定性分析的一个主要问题在于阐释过程中不可避免的主观性。为了尽可能消除“解释者偏见”和主观选择性,定性分析开始遵循严格的程序模板或程序规则,并尝试引入定量分析中的“信度”“效度”“代表性”等概念,通过编码和对比,再加上传统的定性分析标准,如可解释性、透明性和一致性,使得定性研究的过程更加规范、阐释的结果更加客

观,研究的结论更加可信。

5. 研究过程更加有效率:这主要应归功于大量计算机辅助定性数据分析(CAQDA)软件的涌现。从 20 世纪 80 年代以来,定性分析过程的数字化和计算机化,已经是一个不可逆转的大趋势。这种发展趋势与定性研究者的理论取向无关,不管他们的理论立场是实证主义、符号互动论,还是本土方法论,大多数定性研究者都在自己的研究当中,开始采用计算机来辅助定性资料的分析过程。据不完全统计,目前已经有二十多种定性分析的软件,分别隶属于德国、英国、法国、美国等国家。其中,有一些软件是国外研究机构的科研成果,可以免费使用,但比较成熟的定性辅助系统大多是商业软件。这些定性分析的辅助系统,不仅使得研究者从处理大量文字材料的繁复劳动中解放出来,而且能够让研究者共享他们各自分析的细节,从而改变定性研究的流程和研究集体之间的合作方式。同时,由于采用数据库结构,定性资料的管理也更加方便,这就为组织大型定性研究项目(包括多个研究地点、多个研究对象、历时的定性研究)提供了新的可能性。越来越多的定性研究人员开始走出他们的摇椅,坐到计算机屏幕前、湮没在访谈资料和故纸堆中的定性社会学家的形象已经一去不复返了。

6. 定性研究和定量研究的结合更加紧密:在定量分析方法的教材中,定性研究常常被看做是定量研究的前期准备工作,但定性研究者却持完全相反的观点,他们一般认为定性方法是自成一体的,可以完成从形成概念到检验假设的全部研究过程。在实际的应用研究中,定性方法和定量方法常常是交织在一起的,例如,克劳(Currall)等人在研究组织环境重要的群体过程时,通过内容分析把 5 年的参与观察资料量化,然后用统计分析来检验理论假定。格雷(Gray)和邓斯坦(Densten)在研究企业的控制能力时,利用潜变量模型把定性方法和定量方法有机结合在一起。雅各布斯(Jacobs)等人在研究比利时的家庭形态对配偶的家庭劳动分工影响时,首先用定量方法对纵向调查数据进行分析,从定量分析的结果中,又延伸出对核心概念的定性研究。这三个研究分别代表了定量和定性方法相互融合的三个方向:①克劳等人的研究代表着定性方法的实践者试图将定性数据尽可能量化的取向,近年来涌现出的处理调查数据中开放题器的编码问题的工具软件(如 Words at, Smarttext 等,注意:它们都是由著名的统计软件公司出品的处理定性资料的软件),处理定性资料的传统内容分析软件(如 Nvivo、MaxQDA、Kwalitan 等)也开始提供将定性资料转换到常用统计软件的数据接口,这些工具上的革新将加快这种趋势的发展。②格雷和邓斯坦的工作代表了"方法论多元论"的取向,即在应用研究过程中,通过核心概念的测量模型,把定性研究和定量研究结合在一起。③雅各布斯等人的工作则代表了一部分定量研究者对过度形式化的定量方法的不满,并试图通过定性方法加以弥补。在定量研究领域中,对"模型设定"问题的关注,是定量方法重新试图返回定性研究这种取向的另外一种表现。

与社会调查技术和社会研究方法突飞猛进的现实相比,我国学术界在这些方面的论著的出版似乎显得有些迟缓。虽然已经翻译了美国的一小部分经典定量分析教材,如布莱洛克(Blalock)和巴比(Babie)的教材,也有自己编写的一些教材,如袁方等人的《社会研究原理和方法》、卢淑华的《社会统计学》等,此外,偏重软件操作的还有郭志刚的《社会统计分析方法——spss 软件应用》、郭志刚的《logistic 回归模型——方法与应用》、阮桂海的《spss for windows 高级应用教程》等。在《社会学研究》等专业杂志上,也常常有一些定量分析的应用研究,可是专门的方法和应用模型研究却没有,也没有专门的方法研究期刊。仅就定量研究方法的介绍而言,也存在一些缺陷,主要表现在:

1. 原理和操作脱节。

2. 过分依赖某些商业软件，不全面。

3. 与中国的实证研究相脱节。

4. 不能反映当前方法研究的最新进展。

与定量研究方法相比，由于各种原因，定性研究方法的引进和介绍都比较少。在福特基金会资助的方法高级研讨班上，曾讨论过一些定性研究方法。在定性方法研究方面也有少数专著，如袁方和王汉生 1997 年出版的教程，陈向明 2000 年出版的专著。但总体说来，我们对定性研究方法还停留在初步介绍的阶段，主要的介绍也局限在定性研究的研究设计和资料收集的阶段上，对定性分析方法的介绍，则没有能够反映出当代定性方法的最新进展。特别是在定性分析工具（定性分析软件）的引进和研究上，基本上还是一个空白。虽然不乏一些出色的定性研究报告，但从方法研究上讲，我们才刚刚起步。当然，我们同时还应该注意到，在历史学领域，我国对定性资料的鉴别、考据和分析，积累了大量的经验和知识，这也应当是定性方法研究的知识来源之一，应努力发扬光大。

令人欣慰的是，社会研究方法的引进和出版方面相对滞后的状况终于有所改观。重庆大学出版社的编辑，以独到的学术眼光，逆当前出版界唯利是图的不良选题风气，投入了大量的人力、物力，组织出版“万卷方法”。自 2004 年至今，已引进社会科学研究方法方面的专著十余种，在我国社会科学界已经引起了一定的反响。然而，更为可贵的是，重庆大学出版社并未以已经取得的成绩而自满，而是再接再励，在原有“万卷方法”的基础上，进一步组织出版“万卷方法—社会科学研究方法经典译丛”。按我们的设想，“译丛”应该是一个开放的体系，旨在跟踪社会科学研究方法发展的前沿，引进和介绍这一方面的经典著作和最新成果。

“译丛”第一批有《抽样调查设计导论》《社会科学研究设计原理》《社会科学研究测量原理》《社会科学研究分析技术》《问卷设计手册》《回归分析法》《数据再分析法》《抽样调查设计导论》《社会网络分析法》《广义潜变量模型》《定性变量数据分析》和《复杂调查设计和分析方法》（书名也许有变化）等十余种，几乎囊括了研究设计、测量和分析方法的所有领域，涵盖从基础的回归分析到最前沿的潜变量分析和多水平模型等各种分析方法。无论是社会科学各专业的本科生、研究生，还是社会科学研究的学者都将从中有所收获。

“译丛”由中国社会科学院社会学所社会调查与方法研究室的多位研究人员担纲，主译者都是在社会研究方法各个领域中具有相当造诣的教师和研究人员。“译丛”的译者不仅仅把翻译看做是一个“翻译”，而且也把它看做是一次再学习和再创新。

我们期待“译丛”的出版能对社会研究方法的研究、应用和教学有所推动。

沈崇麟　夏传玲

于中国社科院社会学所社会调查与方法研究室

前　言

从 20 世纪 60 年代到现在，有关分类数据的分析方法取得了飞速的发展。本书旨在对这些方法加以介绍，包括那些较早出现的、目前已经被广泛应用的方法。这里，我们尤其强调广义线性模型技术的重要性及其对多元结果变量的扩展，广义线性模型本身则源于对适用于连续变量的线性模型方法的扩展。

目前，由于分类数据分析技术的发展以及分类数据在现实应用中的独特价值，许多统计系或生物统计系都开设了有关分类数据分析的课程。这本书可以用作该类课程的教科书。本书的第 1—7 章涵盖了该类课程的核心内容。其中，第 1—3 章介绍分类结果变量的分布以及传统的二维列联表分析方法。第 4—7 章介绍关于二分和多项分布结果变量的 logistic 回归以及相应的 logit 模型。第 8 章和第 9 章的内容则是用于分析列联表数据的对数线性模型。随着时间的推移，对数线性模型的重要性似乎有所降低，所以本版在一定程度上缩减了对该模型的讨论，并相应增加了有关 logistic 回归的内容。

在过去 10 年间，这一领域的新发展主要集中于对重复测量和其他形式的群组分类数据的分析方法。第 10—13 章讲述这些方法，其中包括边际模型和具有随机效应的广义线性混合模型。第 14—15 章介绍本书所使用的最大似然估计的理论基础以及其他可供选择的估计方法。第 16 章简单回顾了分类数据分析技术的发展历程，并介绍了诸如皮尔逊和费舍尔等著名统计学家的贡献，他们的开创性工作——包括一些激烈的论战——为分类数据分析方法的发展奠定了基础。

本版对第一版的所有章节都做了大量的修改和重写，并扩充了相应内容。本版与前一版的主要区别包括：

- 全新的第 1 章，用来介绍分类数据的分布和统计推断方法。
- 从第 4 章开始一直到本书结尾，系统地将所有模型统一表述为广义线性模型的特例来加以介绍。
- 更加强调关于二分结果变量的 logistic 回归及其对多项分布结果变量的扩展。第 4—7 章着重介绍这些模型，第 10—13 章则将它们扩展到群组数据的情况。
- 增加三个全新的章节用于介绍有关群组数据和相关联的分类数据的分析方法，这些方法在实际应用中正变得越来越重要。
- 增加一个全新的章节介绍这些方法的发展历史。
- 更多的关于“精确”小样本方法以及条件 logistic 回归的讨论。

在本书中，*分类数据分析*（*categorical data analysis*）是指在结果变量为分类变量情况下的数据分析方法。对大多数方法而言，解释变量既可以是定性的，又可以是定量的，如普通的回归分析。因而，本书旨在强调比列联表分析更普遍的分析技术，尽管出于数据表达尽量简单化的考虑，书中大多数例子使用了列联表数据。尽管这些例子往往都比较

简单,但它们可以帮助读者集中精力去理解方法本身,并使读者能够较为方便地利用自己擅长的软件去复制有关结果。

本书的主要特色包括:

- 超过 100 多个“真实”(案例)的数据分析。
- 在各章后面共附有 600 多道习题,其中一部分针对理论和方法,另一部分则强调现实应用中的数据分析。
- 书后的附录给出了如何利用 SAS 软件来使用本书各章所介绍的分析方法。
- 每章后面的注解提供了相应领域的最新进展,以及本书未涉及的许多方法的相应文献。

附录 A 综述了使用本书所介绍的方法需要的统计软件,包括如何利用 SAS 完成书中提到的数据分析,并给出了一个可供参考的网站(www. stata. ufl. edu/ ~ aa/cda/cda. html)。该网站的内容包含:①有关其他软件使用的信息(如 R,S-Plus,SPSS,以及 Stata);②利用 SAS 程序进行相应分析的完整数据;③许多题号为奇数的习题的简略答案;④对本书在早期印刷中所出现错误的更正;⑤更多的习题。作为学习这些方法的辅助材料,我建议读者在阅读本书的同时参阅该附录或者相应的专门手册。

在撰写本书时,我努力尝试使那些来自不同背景、正在学习分类数据分析研究生课程的学生都可以使用。但是,在写作过程中,我还是重点考虑了应用统计学家和生物统计学家的需求。我希望本书能帮助他们了解该领域的最新进展,了解那些在传统统计学教程中未获得足够重视的方法。

新的分析技术的发展促进了各个学科中有关分类结果变量数据的增长,同时这些数据的出现也会促进分析技术的进一步发展。这些学科主要包括社会科学、行为科学和生物医学,同时还包括公共卫生、人类遗传学、生态学、教育学、市场营销,以及工业质量控制等。因此,尽管本书主要面向统计学家和生物统计学家,我也希望以上领域的研究者都能从本书中获益。

本书的读者应当具备一定的回归模型、方差分析以及最大似然估计等统计理论基础。即便不具备太多统计理论背景的读者也应该能够理解本书中有关方法的大部分讨论。略过书中带有星号的章节基本上不会影响对本书的整体阅读。以应用为主要目的的读者,可以略过第 4 章关于广义线性模型理论的大部分内容。不过,这本书与我的《分类数据分析简介》(*An Introduction to Categorical Data Analysis*)(Wiley, 1996)一书相比,在技术层面上要求明显更高,而且内容也更加清楚和完整。

感谢所有对我的书稿提出过宝贵建议或者通过其他形式向我提供帮助的人。我尤其感谢 Bernhard Klingenberg 认真阅读了书中的部分章节并给出了很多有益建议,Yongyi Min 绘制了书中大量的图并提供了软件支持,以及 Brian Caffo 帮助我准备了部分例子。非常感谢参与审阅书稿的 Rosyln Stone 和 Brian Marx,以及对部分章节提出重要建议的 Brian Caffo、I-Ming Liu 和 Yongyi Min。感谢在布朗大学使用本书草稿作为讲义并提出宝贵意见的 Constantine Gatsonis 以及他的学生们。其他曾提出建议或提供其他形式帮助的人包括 Patricia Altham, Wicher Bergsma, Jane Brockmann, Brent Coull, Al DeMaris, Regina Dittrich, Jianping Dong, Herwig Friedl, Ralitza Gueorguieva, James Hobert, Walter Katzenbeisser, Harry Khamis, Svend Kreiner, Joseph Lang, Jason Liao, Mojtaba Ganjali, Jane Pendergast, Michael Radelet, Kenneth Small, Maura Stokes, Tom Ten Have, Rongling Wu。感谢我在各项研究中的合作者允许我在此使用有关文章中的研究成果,尤其是

Brent Coull、Joseph Lang、James Booth、James Hobert、Brian Caffo 以及 Ranjini Natarajan。感谢那些审阅过本书第一版或为本书提供了例子的人,第一版的前言提及了他们的名字。同时感谢 Wiley 的执行主编 Steve Quigley 对本书的长期鼓励和支持。最后,感谢我的妻子 Jacki Levine 在方方面面所给予的持续支持,尤其是长期的写作占据了许多本应属于我们共有的时光。

阿兰·阿格莱斯蒂(ALAN AGRESTI)
美国佛罗里达州盖恩斯维尔市(Gainesville, Florida)
2001 年 11 月

目　录

1 引言:分类数据的分布与统计推断

从评估新的临床治疗方式到探讨影响人们观点和行为的因素,现在的数据分析者正面临着大量关于分类数据分析方法的应用。在本书中,我们介绍这些方法以及它们背后的理论。

与因变量为连续变量的统计方法在20世纪初就取得了长足进展相比,针对分类结果变量的统计方法的成熟则晚得多。尽管在1900年左右英国统计学家卡尔·皮尔逊(Karl Pearson)就在这一领域作了许多重要贡献,直到1960年以前关于分类结果变量的统计模型的发展仍然相当缓慢。在本书中,我们会介绍这些在今天仍具重要地位的早期工作,但是我们的重点将主要是讨论近期发展的模型分析方法。在概述本书所包括的内容之前,让我们首先来介绍一下分类数据的主要类型。

1.1 分类数据

分类变量(*categorical variable*)在测量尺度上是由一系列的类别组成的。比如说,政治态度的测量经常可以划分为激进派、中间派和保守派,X射线对乳腺癌的诊断结果可以用正常、良性、可能良性、疑似以及恶性等类别来表示。

分类数据分析方法的发展是与社会科学和生物医学的研究需求分不开的。在社会科学中,分类尺度在对态度和观点的测量中占据统治地位。在生物医学中,分类尺度的测量也同样常见,比如对一种医学治疗方式是否有效的测量。

尽管分类数据在社会科学和生物医学中非常普遍,对分类数据的应用绝不仅仅局限于这些领域。分类数据经常出现在行为科学(如精神疾病可以划分为精神分裂、抑郁、神经衰弱等类别)、流行病学和公共卫生(如最近一次性行为的避孕方式的类别包括没有、避孕套、避孕药、宫内节育器以及其他)、遗传学(一个后代所继承的等位基因的类型)、动物学(如美洲鳄的主要觅食类型可分为鱼类、无脊椎动物、爬虫动物)、教育学(如学生对一个考试题目的回答结果分为正确与错误),以及市场营销学(如顾客对某种产品的主要品牌的偏好可以划分为品牌A、品牌B和品牌C)。分类数据甚至会出现在高度量化的领域,如工程学和工业质量控制。具体的例子包括按照产品是否符合一定的标准对其进行分类,以及对某些特质的主观评价:某种纤维质感的柔软度,某种食品的口味,或者工人认定完成某项任务的难易程度。

分类变量本身又可以分为许多种类。在这一节,我们讨论对分类变量以及其他变量进行划分的方式。

1.1.1 结果变量与解释变量的区分

大多统计分析都会区分结果变量(或者因变量)和解释变量(或者自变量)。例如,回归模型描述一个结果变量——如住房市场价格——的均值,如何随着解释变量如居住面积和位置的取值的变化而变化。在本书中,我们重点介绍有关结果变量为分类变量的分析方法。当然,与普通回归模型的情形相同,解释变量可以是任何类型的变量。

1.1.2 定类与定序尺度的区分

分类变量包括两种主要的测量尺度。变量的类别间不存在自然顺序的称为*定类*(*nominal*)变量。定类变量的例子包括宗教信仰(类别有天主教徒、新教徒、犹太教徒、穆斯林以及其他),上班的交通方式(汽车、自行车、公车、地铁、步行),喜好的音乐类型(古典、乡村、大众、爵士、摇滚),以及居住选择(公寓、共有住宅单元、独立住房及其他)。对定类变量而言,类别间的排列次序没有意义。因此,统计分析本身也无需考虑排序的问题。

许多分类变量的类别确实具有顺序,这样的变量称为*定序*(*ordinal*)变量。定序变量的例子包括汽车规格(微型、小型、中型、大型),社会阶层(上层、中层、下层),政治倾向(开放、中间、保守),以及病人病情(好、一般、严重、危急)。定序变量的类别之间存在次序,但是各类别之间的间距是不确定的。尽管政治观点为中间派别的人比保守派的人更开放,但没有一个数值可以衡量到底中间派比保守派更开放多少。针对定序变量的统计方法需要考虑有关类别之间次序的信息。

定距变量(*interval variable*)是指任意两个取值之间存在确切的距离的变量。例如,血压水平、电视机的使用寿命、服刑期的长度以及年收入都是定距变量(有一种特殊的定距变量,其取值之间的比也是有意义的,这样的定距变量有时称为*定比变量*(*ratio variable*))。

一个变量的尺度划分取决于它的测量方式。例如,当用来区分公立学校或私立学校时,“教育”仅仅是一个定类变量;当测量的是最高受教育程度时,它是一个定序变量,其类别可包括未接受教育、高中、大学、硕士和博士;当测量受教育年限时,它就成了一个定距变量,其取值为0,1,2,…。

一个变量的测量尺度决定了在分析中应该使用什么样的统计方法。在测量尺度的区分中,定距变量是最高级别的,定序变量其次,定类变量级别最低。针对某一尺度变量的统计方法可以应用于比其测量尺度层级更高的变量,但是不能应用于比其尺度层级更低的变量。例如,针对定类变量的统计方法可以用来分析定序变量,只是忽略了该变量类别间的排序。但是,针对定序变量的统计方法不能应用于分析定类变量,因为后者的类别间不存在有意义的顺序关系。通常来说,在数据分析时,最好应用与实际的测量尺度相符合的分析方法。

由于本书关注分类的结果变量,我们将讨论对定类变量和定序变量的分析。这些方法也适用于只能取少数不同值的定距变量(如结婚的次数)或者可以将其取值分组为有序的类别的定距变量(如受教育年限可以划分为:<10 年、$10\sim12$ 年、>12 年)。

1.1.3 连续变量与离散变量的区分

根据可能取值的数量变量可以划分为*连续*(*continuous*)变量*和离散*(*discrete*)变量。由于受测量工具精确性的限制,在实际测量时,所有的变量都采取离散的形式。在应用

中,这种"连续—离散"的区分取决于变量是可以取很多个值还是较少的值。例如,统计学家常常将可以取很多个值的离散定距变量(如考试成绩)视为连续变量,并利用分析连续变量的方法对其进行分析。

本书主要讨论特定类型的离散结果变量:①定类变量;②定序变量;③可以取相对较少值的离散定距变量;④能够划分为较少组别的连续变量。

1.1.4 定量变量与定性变量的区分

定类变量是*定性的*(*qualitative*)变量——它的不同取值之间存在质的区别,而不是量的差异。定距变量是*定量的*(*quantitative*)变量——其不同取值表示所关注的特质的量的多少。至于定序变量到底属于定量的还是定性的变量并不是很明确。数据分析者经常将其视为定性的变量,并利用定类变量的分析方法来处理定序变量。但是在许多方面,定序变量实际上更接近于定距变量而不是定类变量。定序变量包含了重要的量化特征:每一类别都比其他类别具有*较多*或*较少*的特质。而且,尽管无法测量,一般都认为在定序变量背后隐含着潜在的连续变量。例如,对政治倾向的定序测度(开放、中间、保守),可以看作是对一个内在的连续特质的一种粗糙测量。

数据分析者常常利用定序变量所具有的量化特征,对其类别赋以确定的数值或假定存在一个潜在的连续分布而加以分析。这样做往往需要很强的判断能力以及有经验的研究者的具体指导,不过其好处在于可以借用许多针对定距变量的分析方法。

1.1.5 本书的结构

本书所讨论的关于分类结果变量的模型与用于连续性结果变量的回归模型相似;其主要区别在于,分类结果变量的模型假定二项、多项或者泊松分布而不是正态分布。在这里我们将重点讨论两种模型:logistic 回归和对数线性模型。普通的 *logistic 回归模型*(*logistic regression models*),也称为 *Logit 模型*(*Logit models*),应用于*二分的*(*binary*)结果变量(即具有两个类别),并假定该变量服从二项分布。扩展的 logistic 模型可应用于包括多个类别的结果变量,假定相应变量服从多项分布。*对数线性模型*(*loglinear models*)应用于计数数据(count data),并且假定服从泊松分布。Logistic 回归与对数线性模型之间存在着一定的等价性。

本书可以分为四个主要部分。在第一部分,也即第 1—3 章,我们综述有关单变量和双变量分类数据的描述和统计推断方法,包括离散分布、统计推断方法,以及关联的度量指标。此外,第一部分还回顾了 1960 年之前所发展的不基于模型的分析方法。

在第二部分也即本书的主体部分,第 4—9 章,我们介绍针对分类结果变量的模型。第 4 章介绍一组*广义线性模型*(*generalized linear models*),本书中讨论的所有模型都是广义线性模型的特例。我们重点介绍针对二分结果变量和计数结果变量的模型。第 5 章和第 6 章讨论关于二分结果变量的重要模型,logistic 回归。在第 7 章,我们给出 logistic 模型对多类别的定类和定序结果变量的扩展。第 8 章介绍有关多元分类结果变量的模型分析,并展示如何应用对数线性模型分析相应的列联表计数的关联和交互模式。在第 9 章,我们探讨如何构建对数线性模型和相应的 Logit 模型,以及一些与此有关的其他模型。

在第三部分,即第 10—13 章,我们讨论处理重复测量数据和其他群组数据的模型。第 10 章介绍针对配对数据中分类结果变量的模型。例如,这种模型可用于分析对同一群体在两个不同时点进行了两次测量的数据中的分类结果变量。第 11 章给出更为一般

情况下关于重复测量的分类结果变量的模型。在第 12 章,我们介绍一组非常广泛的模型,即利用随机效应来处理数据中的相依性的*广义线性混合模型*(*generalized linear mixed models*)。第 13 章讨论对第 10—12 章中所介绍的模型的进一步扩展和应用。

本书的最后一部分比较偏重于理论。第 14 章推导分类数据模型的渐近理论。该理论是在大样本情况下模型参数估计值和拟合优度统计量所具有的特性的基础。在第 14 章以及本书的其他地方,我们主要讨论最大似然估计。第 15 章则主要介绍其他的估计方法,如贝叶斯估计。第 16 章相对独立,它回顾了分类数据分析方法的发展历程。

多数分类数据的分析方法都要求大量的运算,因而统计软件对有效利用这些方法必不可少。在附录 A 我们介绍可以实现本书中所讨论的分析方法的软件,并且示范如何应用 SAS 处理书中所提到的例子。读者可以通过以下网站 *www. stat. ufl. edu/ ~ aa/cda/cda. html* 来下载相应的程序和数据,或是了解其他软件的有关信息。

第 1 章介绍的是一些统计基础。在第 1.2 节中,我们回顾关于分类数据的主要分布类型:二项分布、多项分布,以及泊松分布。第 1.3 节回顾运用最大似然法进行统计推断的基本原理。在第 1.4 和 1.5 节中,我们展示如何对二项分布和多项分布的参数进行显著性检验并构建置信区间。

1.2 分类数据的分布

统计推断的数据分析要求对生成数据的随机过程进行一定的假设。对结果变量为连续变量的回归模型而言,正态分布假设起着核心的作用。在这一节,我们介绍分类结果变量的三种主要分布:*二项*(*binomial*)分布、*多项*(*multinomial*)分布,以及*泊松*(*Poisson*)分布。

1.2.1 二项分布

许多应用是针对二分观测值的总数固定为 n 个的情况。令 $y_1, y_2, \cdots, y_n$ 分别表示 n 次独立同分布的试验结果,$P(Y_i=1)=\pi$,$P(Y_i=0)=1-\pi$。通常,我们用 1 和 0 分别表示“成功”和“失败”的结果。*同分布试验*(*identical trials*)意味着对每一次试验而言,成功的概率 π 都相等。*独立试验*(*independent trial*)则表示 $\{Y_i\}$ 是独立的随机变量。这种情况通常称为*伯努利试验*(*Bernoulli trials*)。总的成功次数,$Y=\sum_{i=1}^{n} Y_i$,服从基数为 n、参数为 π 的*二项分布*(*binomial distribution*),表示为 $\text{bin}(n,\pi)$。

对于 Y 的可能结果 y,其概率密度函数为

$$p(y) = \binom{n}{y}\pi^{y}(1-\pi)^{n-y}, \qquad y = 0,1,2,\cdots,n, \tag{1.1}$$

这里二项系数 $\binom{n}{y}=n!/[y!\,(n-y)!]$。由于 $E(Y_i)=E(Y_i^2)=1\times\pi+0\times(1-\pi)=\pi$,

$$E(Y_i)=\pi \quad 并且 \quad \text{var}(Y_i)=\pi(1-\pi)。$$

$Y=\sum_i Y_i$ 所服从的二项分布的均值和方差分别为

$$\mu = E(Y) = n\pi \quad 以及 \quad \sigma^2 = \text{var}(Y) = n\pi(1-\pi)。$$

二项分布的偏度(skewness)由 $E(Y-\mu)^3/\sigma^3=(1-2\pi)/\sqrt{n\pi(1-\pi)}$ 来表示。对于固定的 π,随着 n 的增加,二项分布收敛于正态分布。

相继进行的二分观测并不一定就是相互独立的或服从同分布的。因而,有时候我们

需要应用其他的分布。一种情况是从一个有限总体中无放回地抽取二分结果,比如从由20 个学生组成的班中抽取 10 人得到的性别分布。这时,就要用到在本书第 3.5.1 节中介绍的*超几何分布*(*hypergeometric distribution*)。在第 1.2.4 节,我们会提及另一种不满足这些二项分布假定的情形。

1.2.2 多项分布

一些试验的可能结果多于两种。假定在 n 次独立同分布的试验中,每一次的结果可以是 c 种可能类别中的某一类。如果第 i 次试验的结果为类别 j,则令 $y_{ij}=1$,否否则 $y_{ij}=0$。那么 $\boldsymbol{y}_i=(y_{i1},y_{i2},\cdots,y_{ic})$ 就表示一个多项分布试验,并且 $\sum_j y_{ij}=1$;例如,(0,0,1,0) 表示在四种可能类别中结果是第三种的情况。注意 y_{ic} 是冗余的,它的值完全可以用其他类别的取值来线性表示。令 $n_j=\sum_i y_{ij}$ 表示试验中出现第 j 种结果的次数,那么计数$(n_1,n_2,\cdots,n_c)$服从*多项分布*(*multinomial distribution*)。

令 $\pi_j=P(Y_{ij}=1)$ 代表每一次试验的结果落在第 j 类的概率。多项分布的概率密度函数可表示为

$$p(n_1,n_2,\cdots,n_{c-1})=\left(\frac{n!}{n_1!n_2!\cdots n_c!}\right)\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_c^{n_c}。\tag{1.2}$$

由于 $\sum_j n_j=n$,所以多项分布只包含$(c-1)$个维度,并且 $n_c=n-(n_1+\cdots+n_{c-1})$。二项分布可以看作是多项分布在 $c=2$ 时的一个特例。

对于多项分布而言,

$$E(n_j)=n\pi_j,\quad \text{var}(n_j)=n\pi_j(1-\pi_j),\quad \text{cov}(n_j,n_k)=-n\pi_j\pi_k。\tag{1.3}$$

关于协方差的推导过程,详见第 14.1.4 节。在多项分布中,每个 n_j 的边际分布都服从二项分布。

1.2.3 泊松分布

有时,计数数据(count data)并不是来自于总数固定的试验。例如,如果 y 等于下一周意大利高速公路上所发生的车祸导致的死亡人数,这时对于 y 而言,就不存在一个固定的上限 n(如果你在意大利开过车,这一点便很清楚)。由于 y 必须是一个非负的整数,它的分布应当落在非负整数的范围内。满足这样条件的最简单的分布就是*泊松*(*Poisson*)分布。它的概率仅取决于一个参数,即均值 μ。泊松分布的概率密度函数(Poisson, 1837, p. 206)为

$$p(y)=\frac{e^{-\mu}\mu^y}{y!},\quad y=0,1,2,\cdots。\tag{1.4}$$

泊松分布满足 $E(Y)=\text{var}(Y)=\mu$。另外,它的众数是唯一的,等于 μ 的整数部分。泊松分布的偏度可以表示为 $E(Y-\mu)^3/\sigma^3=1/\sqrt{\mu}$。随着 μ 的增大,泊松分布趋近于正态分布。

泊松分布可用于分析在一定时间或空间内某随机事件发生的次数,这时,在互不交叉的时间段或空间区域内,不同结果发生的可能性是相互独立的。当二项分布中的 n 很大并且 π 很小时,可以用泊松分布作为对二项分布的一种近似,其中 $\mu=n\pi$。因此,假设下周在意大利开车的五千万人均可被视为相互独立的试验,且该周每人死于严重交通事故的概率均为0.000 002,那么下周所发生的死亡总数 Y 服从 bin(50 000 000,0.000 002)分布,或者近似服从 $\mu=n\pi=50\,000\,000(0.000\,002)=100$ 的泊松分布。

泊松分布的一个重要特征是其方差等于均值。当泊松分布的均值较大时,样本计数的变动范围也较大。因此,当每周平均致命事故的数量为 100 时,事故数量的变动程度要比每周平均事故数为 10 时大得多。

1.2.4 过度离散

在实际应用中,计数变量常常显示出比二项分布或泊松分布所预测的方差更大的变动性。这种现象被称为*过度离散*(*overdispersion*)。如在前文中,我们假定在下周每个人死于严重交通事故的概率是相等的。而现实的情况往往是,由于开车的总时长、是否使用安全带,以及地理位置等因素的影响,死亡概率会存在差异。这种差异性导致死亡事故发生次数的变动性比泊松分布所预测的要大。

假设在 μ 给定的情况下 Y 是一个方差为 $\mathrm{var}(Y|\mu)$ 的随机变量,但是由于如上文中所提及的没有测量的因素,μ 本身可能变动。令 $\theta = E(\mu)$,那么无条件地,

$$E(Y) = E[E(Y \mid \mu)], \quad \mathrm{var}(Y) = E[\mathrm{var}(Y \mid \mu)] + \mathrm{var}[E(Y \mid \mu)]。$$

当 Y 在给定 μ 的条件下服从泊松分布时,那么存在 $E(Y) = E(\mu) = \theta$ 并且 $\mathrm{var}(Y) = E(\mu) + \mathrm{var}(\mu) = \theta + \mathrm{var}(\mu) > \theta$。

由于存在导致过度离散的因素,假定计数变量服从泊松分布通常都过于简化了。当计数变量的方差超过均值时,*负二项*(*negative binomial*)分布可以处理此类情况。关于负二项分布的详细介绍,参见第 4.3.4 节。

数据分析中关于二项(或多项)分布的假定有时也会由于变量过度离散而不切实际。当真实的分布是由多个不同的二项分布所形成的混合体,并且二项分布参数由于未测量的变量而变动时,就会出现过度离散问题。举例来说,假定在实验中将受孕的母鼠暴露到一种毒素中,并在一周之后观察每只老鼠体内的胚胎出现畸形的数量。令 n_i 表示第 i 只老鼠体内所含的胚胎数。母鼠本身也会由于一些其他未测量的因素而存在差别,比如它们的体重、总体健康状况,以及基因构成。这种母鼠之间的差别会导致出现畸形的概率 π 的额外变动。因此,每只老鼠体内出现畸形胚胎的数量的分布可能会集中在 0 值或者集中在 n_i,表现出比依据一个单一的 π 值进行的二项分布抽样更离散的趋势。当 π 在同一母鼠的不同胚胎之间按照某种分布存在变动时(习题 1.12),也会出现过度离散的问题。在第 4 章、第 12 章,以及第 13 章,我们将介绍处理与二项分布和泊松分布相比数据存在过度离散问题的方法。

1.2.5 泊松分布与多项分布的联系

假设在接下来一周,y_1 = 意大利死于汽车事故的人数,y_2 = 死于飞机事故的人数,并且 y_3 = 死于火车事故的人数。关于 (Y_1, Y_2, Y_3) 的泊松模型将这些变量视为相互独立的泊松随机变量,其参数为 (μ_1, μ_2, μ_3)。那么,$\{Y_i\}$ 的联合分布概率密度函数等于形式为式 1.4 的三个概率密度函数的乘积。总的死亡人数 $n = \sum Y_i$ 也服从一个参数为 $\sum \mu_i$ 的泊松分布。

按照泊松抽样,总计数 n 是随机的而不是固定的。如果我们假定一个 n 为确定值的泊松模型,$\{Y_i\}$ 就不再服从泊松分布,因为每一个 Y_i 都不能大于 n。给定 n,$\{Y_i\}$ 也不再相互独立,原因是一个变量的取值会影响其他变量取值的可能范围。

对于 c 个相互独立的泊松变量,并且 $E(Y_i) = \mu_i$,我们可以推导出它们在给定 $\sum Y_i = n$ 的限定下的条件分布。满足这一条件的一组计数 $\{n_i\}$ 的条件概率为

$$P[(Y_1=n_1,Y_2=n_2,\cdots,Y_c=n_c)\mid \sum Y_j=n]$$
$$=\frac{P(Y_1=n_1,Y_2=n_2,\cdots,Y_c=n_c)}{P(\sum Y_j=n)}$$
$$=\frac{\prod_i[\exp(-\mu_i)\mu_i^{n_i}/n_i!]}{\exp(-\sum\mu_j)(\sum\mu_j)^n/n!}=\frac{n!}{\prod_i n_i!}\prod_i\pi_i^{n_i},\qquad(1.5)$$

其中$\{\pi_i=\mu_i/(\sum\mu_j)\}$。这正好就是参数为$(n,\{\pi_i\})$的多项分布,其特征由样本规模$n$以及概率$\{\pi_i\}$决定。

许多分类数据的分析都基于多项分布的假定。这样的分析通常会与基于泊松分布假定的分析具有相同的参数估计值,因为这两种分布的似然函数很相似。

1.3 分类数据的统计推断

对结果变量分布的选择仅仅是数据分析中的一步。在实际应用中,所选取的分布对应的参数值也是未知的。在这一节我们回顾利用样本数据对参数进行统计推断的方法。在接下来的第 1.4 节和 1.5 节将分别介绍有关二项分布和多项分布参数的情况。

1.3.1 似然函数和最大似然估计

在本书中我们通过*最大似然法*(*maximum likelihood*)进行参数估计。在弱规则性条件下,如参数空间具有固定的维度且其真值落在空间内部,最大似然估计值具有一些很好的特性:它们服从大样本正态分布;具有渐近一致性,即随着n的增加收敛于参数的真值;并且具有渐近有效性,即在大样本情况下其标准误不大于其他估计方法的标准误。

给定数据,对于某个选定的概率分布,*似然函数*(*likelihood function*)就是观察到这些数据的概率,可以表示为关于未知参数的函数。最大似然(ML)估计是使似然函数取最大值时的参数值。这也是使所观察到的数据以最大的概率发生的参数值。最大化似然函数的参数值同时也是使似然函数的对数取最大值的参数值。最大化对数似然函数可以将似然函数的乘积运算转化为求和运算,因而更为简便。

一般来说我们用β表示某个参数,$\hat\beta$表示相应的最大似然估计。似然函数由$l(\beta)$来表示,而对数似然函数则表示为$L(\beta)=\log[l(\beta)]$。对很多模型来说,$L(\beta)$是一个凹函数,$\hat\beta$就是令对数似然函数的导数为 0 的点。因而,最大似然估计就是似然方程$\partial L(\beta)/\partial\beta=0$的解。通常来说,$\beta$往往是多维的,表示为向量$\boldsymbol\beta$,而$\hat{\boldsymbol\beta}$则是一组似然方程的解。

令 SE 表示$\hat\beta$的标准误,并令$\mathrm{cov}(\hat{\boldsymbol\beta})$表示$\hat{\boldsymbol\beta}$的渐近协方差矩阵。在规则性条件下(Rao, 1937, p. 364),$\mathrm{cov}(\hat{\boldsymbol\beta})$是*信息矩阵*(*information matrix*)的逆矩阵。其中,信息矩阵中的第(j,k)个元素为

$$-E\left(\frac{\partial^2L(\boldsymbol\beta)}{\partial\beta_j\partial\beta_k}\right)。\qquad(1.6)$$

标准误等于信息矩阵的逆矩阵中对角线元素的平方根。对数似然函数的曲率(curvature)越大,标准误就越小。这是因为,曲率越大意味着对数似然值会随着$\boldsymbol\beta$偏离$\hat{\boldsymbol\beta}$而迅速下降;进而,如果$\boldsymbol\beta$的取值接近而不是远离$\hat{\boldsymbol\beta}$,那么所观察到的数据确实发生的可能性要大得多。

1.3.2 二项分布参数的似然函数和最大似然估计

似然函数中涉及参数的那一部分被称作核(*kernel*)函数。由于似然函数的最大化是针对参数而言的,似然函数中除核函数以外的部分与最大化运算无关。

具体而言,考虑二项分布的情况(式 1.1)。二项系数$\binom{n}{y}$本身对似然函数最大化时 π 的相应取值没有影响。因而,我们可以忽略掉这一部分,而把核函数当做似然函数进行最大化。这时,二项分布的对数似然函数就成了

$$L(\pi) = \log[\pi^y(1-\pi)^{n-y}] = y\log(\pi) + (n-y)\log(1-\pi)。\tag{1.7}$$

将该函数对 π 求导可得

$$\partial L(\pi)/\partial\pi = y/\pi - (n-y)/(1-\pi) = (y-n\pi)/\pi(1-\pi)。\tag{1.8}$$

令上式等于 0 便得出了似然方程,它的解为 $\hat{\pi}=y/n$,即在 n 次试验中成功次数所占的样本比例。

计算$\partial^2L(\pi)/\partial\pi^2$ 的期望值,合并同类项,可得

$$-E[\partial^2L(\pi)/\partial\pi^2] = E[y/\pi^2 + (n-y)/(1-\pi)^2] = n/[\pi(1-\pi)]。\tag{1.9}$$

因此,$\hat{\pi}$ 的渐近方差为 $\pi(1-\pi)/n$。这很容易理解。由于 $E(Y)=n\pi$, $\mathrm{var}(Y)=n\pi\cdot(1-\pi)$,因而 $\hat{\pi}=Y/n$ 的分布具有以下均值和标准误:

$$E(\hat{\pi}) = \pi,\qquad \sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}。$$

1.3.3 沃尔德检验、似然比检验和计分检验

通过似然函数进行大样本统计推断有三种标准的方法。我们先介绍如何使用这些方法对零假设 $H_0:\beta=\beta_0$进行显著性检验,然后讨论假设检验与区间估计的关系。这些方法都利用了最大似然估计值的大样本正态分布特性。

在 $\hat{\beta}$ 的标准误(SE)非零的情况下,检验统计量

$$z = (\hat{\beta}-\beta_0)/\mathrm{SE}$$

在当 $\beta=\beta_0$ 时近似服从标准正态分布。我们可以通过标准正态分布表查找关于 z 的单边或双边 P 值。同样地,作为双边检验的另一种方法,z^2 服从具有 1 个自由度(df)的卡方零分布;这时检验的 P 值就是在卡方右尾取值大于观察值的概率。这种利用非零标准误的统计量,被称为*沃尔德统计量*(*Wald Statistic*)(Wald, 1943)。

多元形式的沃尔德检验 $H_0:\boldsymbol{\beta}=\boldsymbol{\beta}_0$具有以下检验统计量

$$W = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'[\mathrm{cov}(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)。$$

在向量或矩阵右上角的斜撇表示该向量或矩阵的转置。非零的协方差是基于在 $\hat{\boldsymbol{\beta}}$ 处对数似然函数的曲率(式 1.6)。由于 $\hat{\boldsymbol{\beta}}$ 渐近服从多元正态分布,W 渐近服从卡方分布。其自由度(df)等于 $\mathrm{cov}(\hat{\boldsymbol{\beta}})$的秩,即 $\boldsymbol{\beta}$ 中非冗余(nonredundant)参数的个数。

第二种通用的方法利用似然函数的两个最大值之比:①在 H_0 成立的情况下所有可能参数取值中最大的似然函数值;②在所有满足 H_0 或者备择假设 H_a 的更大范围的参数可能取值中最大的似然函数值。令 l_0 代表在 H_0 成立时似然函数的最大值,l_1 代表在更一般的条件下(即 $H_0\cup H_a$ 成立时)似然函数的最大值。例如,对于参数向量 $\boldsymbol{\beta}=(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1)'$以及 $H_0:\boldsymbol{\beta}_0=0$,$l_1$ 是根据$\boldsymbol{\beta}$ 的取值内数据最可能发生的情况下所计算的似然函数;l_0 是在$\boldsymbol{\beta}_1$ 的取值内数据最可能发生的情况下所计算的似然函数,这时$\boldsymbol{\beta}_0=0$。由于 l_0 是在

一组较小参数取值范围内所获得的最大化结果,因此 l_1 总是至少和 l_0 一样大。

两个最大似然值之比 $\Lambda = l_0/l_1$ 的值不大于1。Wilks(1935,1938)表明,$-2\log\Lambda$ 在 $n\to\infty$ 时的极限服从卡方分布,其自由度(df)等于在 $H_0\cup H_a$ 条件下与在 H_0 条件下的参数空间维度之差。*似然比检验统计量*(*likelihood-ratio test statistic*)等于

$$-2\log\Lambda = -2\log(l_0/l_1) = -2(L_0 - L_1),$$

其中 L_0 和 L_1 分别代表相应的对数似然函数的最大值。

第三种方法应用的是*计分统计量*(*score statistic*),是由 R. A. Fisher 和 C. R. Rao 提出来的。计分检验的原理基于对数似然函数 $L(\beta)$ 在零假设值 β_0 时的斜率和曲率。它利用的是*计分函数*(*score function*)

$$u(\beta) = \partial L(\beta)/\partial\beta$$

在 β_0 处的取值。$u(\beta_0)$ 的绝对值一般在 $\hat{\beta}$ 远离 β_0 时变得较大。令 $\iota(\beta_0)$ 表示在 β_0 时 $-E[\partial^2 L(\beta)/\partial\beta^2]$(即信息函数)的取值。计分统计量就是 $u(\beta_0)$ 与它的零假设标准误 $[\iota(\beta_0)]^{1/2}$ 之比。它近似服从标准正态零分布。计分统计量的卡方形式为

$$\frac{[u(\beta_0)]^2}{\iota(\beta_0)} = \frac{[\partial L(\beta)/\partial\beta_0]^2}{-E[\partial^2 L(\beta)/\partial\beta_0^2]},$$

其中的偏导数是指函数对 β 的导数在 β_0 处的取值。在多维参数的情况下,计分统计量是由对数似然函数对 $\boldsymbol{\beta}$ 的偏导数与逆信息矩阵所形成的二次项在 H_0 情况下的取值(即假定 $\boldsymbol{\beta}=\boldsymbol{\beta}_0$)。

图1.1展示了在单一参数情况下,利用对数似然函数 $L(\beta)$ 对 $H_0:\beta=0$ 的三种检验。沃尔德检验利用 $L(\beta)$ 在最大似然估计值 $\hat{\beta}$ 时的特性,即 $(\hat{\beta}/\mathrm{SE})^2$ 服从卡方分布。$\hat{\beta}$ 的标准误(SE)取决于 $L(\beta)$ 在 $\beta=0$ 时的曲率。计分检验取决于 $L(\beta)$ 在 $\beta=0$ 时的斜率和曲率。似然比检验联合使用了在 $\hat{\beta}$ 和 $\beta=0$ 时的 $L(\beta)$ 的信息。它通过卡方统计量 $-2(L_0-L_1)$ 比较 $\hat{\beta}$ 对应的对数似然值 L_1 和 $\beta=0$ 对应的值 L_0 的大小。在图1.1中,似然比检验统计量等于 $\hat{\beta}$ 和 $\beta=0$ 对应的 $L(\beta)$ 之间的垂直距离的两倍。在一定意义上,似然比检验统计量在三种统计量中应用的信息最多,因而也最通用。

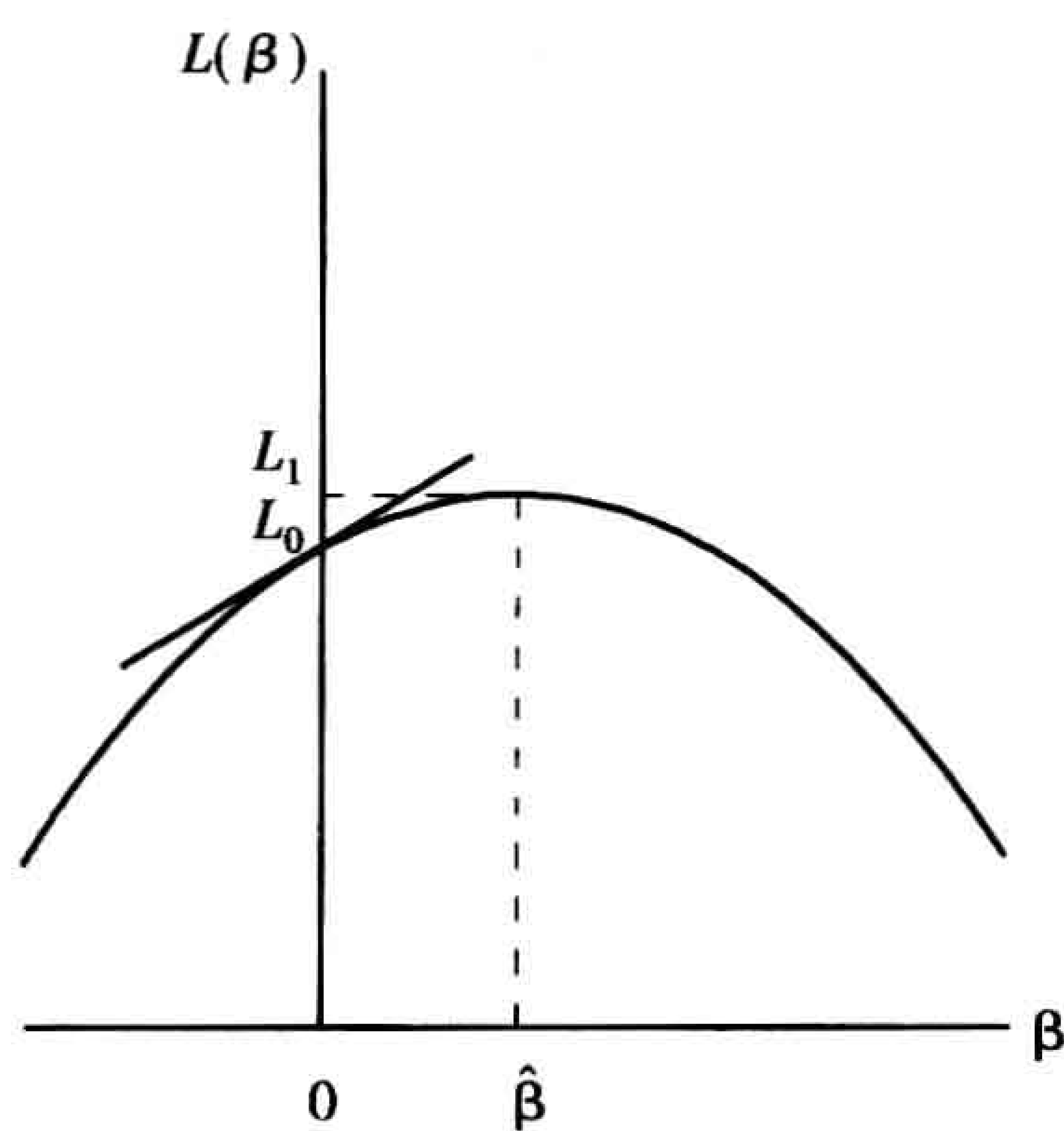


图1.1 对数似然函数以及在对 $H_0:\beta=0$ 进行三种检验时所使用的信息

当 $n\to\infty$ 时,沃尔德检验、似然比检验和计分检验在一定程度上渐近等价(Cox and Hinkley, 1974, Sec. 9.3)。在小样本以及中等规模样本的情况下,似然比检验往往比沃尔德检验更可靠。

1.3.4 构建置信区间

现实应用中,构建参数的置信区间能够提供比对参数取值进行假设检验更丰富的信息。对于上面提到的三种检验方法中的任何一种,都可以通过对该检验的逆运算求得相应的置信区间。例如,关于 β 的95%的置信区间就是检验 $H_0:\beta=\beta_0$ 中 P 值超过0.05的 β_0 的取值范围。

令 z_a 表示标准正态分布中右尾概率为 a 时的 z 值;这是该分布的第 $100(1-a)$ 百分位点。令 $\chi^2_{\mathrm{df}}(a)$ 表示自由度为df的卡方分布的第 $100(1-a)$ 百分位点。基于渐近正态特性的 $100(1-\alpha)\%$ 的置信区间所使用的是 $z_{\alpha/2}$,如95%的置信区间为 $z_{0.025}=1.96$。沃

尔德置信区间就是满足 $|\hat{\beta}-\beta_0|/\text{SE}<z_{\alpha/2}$ 的一组 β_0，由此给出的置信区间为 $\hat{\beta}\pm z_{\alpha/2}(\text{SE})$。基于似然比检验的置信区间就是满足 $-2[L(\beta_0)-L(\hat{\beta})]<\chi_1^2(\alpha)$ 的一组 β_0，如前所述，$\chi_1^2(\alpha)=z_{\alpha/2}^2$。

当 $\hat{\beta}$ 服从正态分布时，对数似然函数具有抛物线形状（即，二阶多项式）。对于小样本的分类数据而言，$\hat{\beta}$ 可能与正态分布相去甚远，因此对数似然函数的形状也与对称的抛物线曲线相去甚远。在中等规模或大样本的情形下，当模型所包含的参数过多时，也可能会发生这种情况。如果出现上述情况，基于 $\hat{\beta}$ 渐近服从正态分布的统计推断就可能存在问题。当沃尔德和似然比的统计推断结果相差很大时，表明 $\hat{\beta}$ 的分布有可能远远不同于正态分布。例如，在 1.4.3 节的例子中，不同方法给出的置信区间相差很大。这时往往可以考虑使用精确小样本分布或者改进简单正态分布假定后的“高阶”渐近方法去进行统计推断（如 Pierce and Peters, 1992）。

在实际应用中，沃尔德置信区间的应用最为广泛，这是因为沃尔德置信区间可以很容易地通过统计软件所报告的最大似然估计值和标准误来构建。现在，越来越多的统计软件开始给出基于似然比的置信区间，而当分析小样本或中等样本的分类数据时，似然比置信区间会更合适。对于最为常见的统计模型——结果变量服从正态分布的回归而言，这三种统计推断的结果完全一致。

1.4 二项分布参数的统计推断

在本节，我们具体讨论对分类数据进行统计推断的方法，给出关于在 n 次独立试验中成功次数为 y 的二项分布参数 π 的统计检验和置信区间。在第 1.3.2 节中我们已经介绍了 π 的似然函数及其最大似然估计值 $\hat{\pi}=y/n$。

1.4.1 对二项分布参数的检验

考虑 H_0: $\pi=\pi_0$。由于 H_0 只包含一个参数，我们使用正态而不是卡方形式的沃尔德和计分检验统计量。这类检验既可以是单边检验也可以是双边检验。沃尔德统计量可表示为

$$z_W=\frac{\hat{\pi}-\pi_0}{\text{SE}}=\frac{\hat{\pi}-\pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}\quad。\tag{1.10}$$

将 π_0 的值代入二项分布的计分统计量（式 1.8）和信息函数（式 1.9），我们得到

$$u(\pi_0)=\frac{y}{\pi_0}-\frac{n-y}{1-\pi_0},\qquad \iota(\pi_0)=\frac{n}{\pi_0(1-\pi_0)}。$$

正态形式的计分统计量简化为

$$z_S=\frac{u(\pi_0)}{[\iota(\pi_0)]^{1/2}}=\frac{y-n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}=\frac{\hat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}。\tag{1.11}$$

尽管沃尔德统计量 z_W 使用 $\hat{\pi}$ 所对应的标准误，而计分统计量 z_S 却使用 π_0 对应的标准误。由于计分统计量应用了零假设的标准误而不是标准误的估计值，因而相比之下计分统计量更好。计分统计量的样本零分布比沃尔德统计量更接近于标准正态分布。

在满足 H_0 的情况下，二项分布的对数似然函数（式 1.7）等于 $L_0=y\log\pi_0+(n-y)\cdot\log(1-\pi_0)$，在更一般的情况下，相应对数似然函数等于 $L_1=y\log\hat{\pi}+(n-y)\log(1-\hat{\pi})$。

这样,似然比检验统计量可简化为

$$-2(L_0 - L_1) = 2\left(y \log \frac{\hat{\pi}}{\pi_0} + (n - y)\log \frac{1 - \hat{\pi}}{1 - \pi_0}\right)。$$

上式可表示为

$$-2(L_0 - L_1) = 2\left(y \log \frac{y}{n\pi_0} + (n - y)\log \frac{n - y}{n - n\pi_0}\right),$$

即,似然比检验统计量将所观察到的成功和失败次数与由零假设模型所拟合的次数进行比较

$$2\sum \text{观察值} \log \frac{\text{观察值}}{\text{拟合值}}。\tag{1.12}$$

在本书后面的内容,我们将看到这个公式对于泊松分布和多项分布的参数检验同样成立。由于在 H_0 的情况下不存在未知参数而在 H_a 的情况下存在一个未知参数,公式 1.12 渐近服从 df = 1 的卡方分布。

1.4.2 二项分布参数的置信区间

显著性检验仅仅表明某个特定的 π 值(如 $\pi = 0.5$)是否可能。我们可以通过置信区间获得参数的可能取值范围的更多信息。

对沃尔德检验统计量进行转换可以得出满足 $|z_W| < z_{\alpha/2}$ 的 π_0 的取值区间,也即

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}。\tag{1.13}$$

这是最早出现的关于参数的置信区间之一(Laplace, 1812:283)。遗憾的是,除非 n 非常大,否则该置信区间很不准确(如 Brown et al., 2001)。它所涵盖的概率通常会低于名义上的置信水平,尤其是当 π 的值接近于 0 或 1 时。对该置信区间进行简单调整,在应用该公式之前在样本中加入 $\frac{1}{2}z_{\alpha/2}^2$ 的各类观测值,所得到的效果就要好得多(习题 1.24)。

计分置信区间是满足 $|z_S| < z_{\alpha/2}$ 的 π_0 值,其端点由下列方程中关于 π_0 的解确定,

$$(\hat{\pi} - \pi_0)/\sqrt{\pi_0(1 - \pi_0)/n} = \pm z_{\alpha/2}。$$

这是关于 π_0 的二次方程式。E. B. Wilson(1927)最先对此进行了探讨,相应置信区间可表示为

$$\hat{\pi}\left(\frac{n}{n + z_{\alpha/2}^2}\right) + \frac{1}{2}\left(\frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2}\right)$$

$$\pm z_{\alpha/2}\sqrt{\frac{1}{n + z_{\alpha/2}^2}\left[\hat{\pi}(1 - \hat{\pi})\left(\frac{n}{n + z_{\alpha/2}^2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2}\right)\right]}。$$

区间的中点 $\tilde{\pi}$ 是关于 $\hat{\pi}$ 和 $\frac{1}{2}$ 的加权平均数,其中权数 $n/(n + z_{\alpha/2}^2)$ 在 $\hat{\pi}$ 给定的情况下随着 n 的增加而增加。合并同类项,区间的中点等于 $\tilde{\pi} = (y + z_{\alpha/2}^2/2)/(n + z_{\alpha/2}^2)$。这是在增加了 $z_{\alpha/2}^2$ 个观测值(每类各半)后所得到的新样本的样本比例。在公式中,$z_{\alpha/2}$ 的系数的平方等于 $\pi = \hat{\pi}$ 和 $\pi = \frac{1}{2}$ 时样本比例的方差的加权平均数,在计算中用调整后的样本规模 $n + z_{\alpha/2}^2$ 来代替 n。这个置信区间比沃尔德置信区间精确得多。

基于似然比的置信区间在计算上更加复杂,但是其原理相对简单。它就是满足似然比检验的 P 值大于 α 的所有 π_0 的取值。同时,它也是使相应的对数似然值与最大似然估计 $\hat{\pi} = y/n$ 对应的对数似然值之差的两倍仍小于 $\chi_1^2(\alpha)$ 的一组 π_0。

1.4.3 例子:素食者的比例

最近为了给一门初级统计学的课程收集数据,我向学生们发了一份问卷,其中一个问题问每个学生是不是素食主义者。在总共 $n=25$ 名学生中,$y=0$ 人回答了"是"。参与调查的学生不是任何一个特定总体的随机样本,但是我们应用这些数据来展示如何计算关于二项分布参数 π 的 95% 置信区间。

由于 $y=0$,$\hat{\pi}=0/25=0$。利用沃尔德方法,关于 π 的 95% 置信区间等于

$$0 \pm 1.96\sqrt{(0.0 \times 1.0)/25}\text{, 也即 } (0,0)\text{。}$$

当观察值落在样本空间的边界时,沃尔德方法常常无法给出有意义的答案。

相反,由计分方法计算的 95% 置信区间等于(0.0,0.133)。这个区间显然更为可信。举例来说,对于 H_0: $\pi=0.5$,相应的计分检验统计量为 $z_S=(0-0.5)/\sqrt{(0.5\times0.5)/25}=-5.0$,因此 0.5 没有落在该置信区间内。相反,对于 H_0: $\pi=0.10$,$z_S=(0-0.10)/\sqrt{(0.10\times0.90)/25}=-1.67$,所以 0.10 落在了该置信区间里面。

当 $y=0$ 且 $n=25$ 时,似然函数的核函数是 $l(\pi)=\pi^0(1-\pi)^{25}=(1-\pi)^{25}$。此时对数似然函数(式 1.7)可简化为 $L(\pi)=25\log(1-\pi)$。注意 $L(\hat{\pi})=L(0)=0$。似然比的 95% 置信区间就是使似然比统计量满足以下条件的一组 π_0:

$$\begin{aligned}-2(L_0-L_1) &= -2[L(\pi_0)-L(\hat{\pi})] \\ &= -50\log(1-\pi_0) \leqslant \chi_1^2(0.05)=3.84\text{。}\end{aligned}$$

区间的上限为 $1-\exp(-3.84/50)=0.074$,即该置信区间等于(0.0,0.074)(在本书中,我们统一使用自然对数,因而它的反函数为指数函数 $\exp(x)=e^x$)。图 1.2 显示了似然函数、对数似然函数以及相应的关于 π 的置信区间。

这三种大样本方法得出的结论之间差别很大。当 π 的值接近于 0 时,$\hat{\pi}$ 的样本分布在小样本的情况下会严重向右偏斜。因而,这时有必要考虑不采用渐近近似的其他方法。

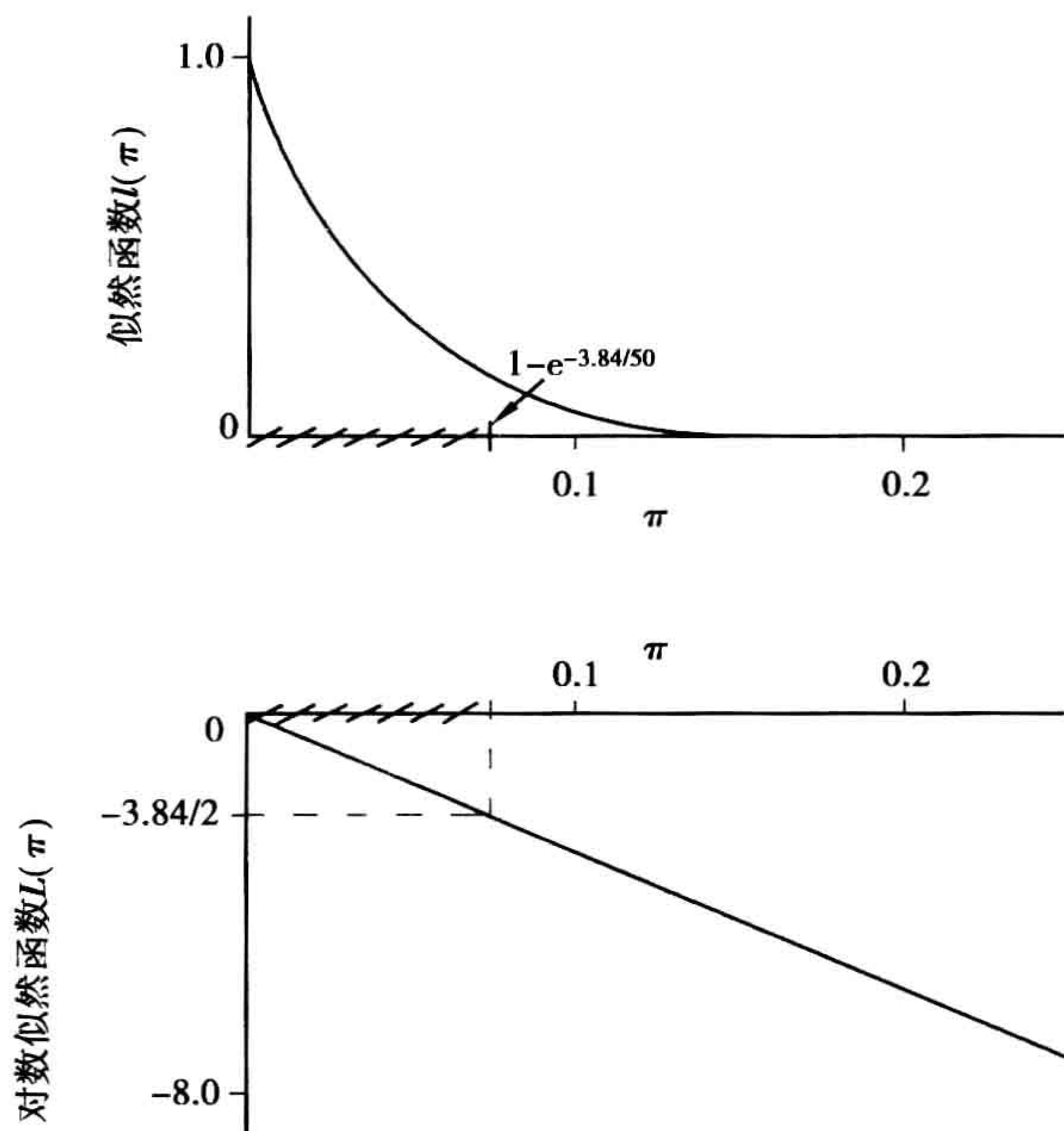


图 1.2　在 $n=25$ 的试验中 $y=0$ 时的二项分布似然函数、对数似然函数,以及 π 的置信区间

1.4.4 精确小样本统计推断*[①]

在现有的运算能力下,对于统计量如 $\hat{\pi}$ 的分布,没必要依赖于大样本的近似。我们可以直接使用二项分布而不是近似的正态分布对其进行统计检验和构建置信区间。这种统计推断方法对小样本而言是一种自然的选择,但是该方法同样适用于其他任何规模的样本。

我们通过关于素食主义的调查数据来介绍如何对 $H_0: \pi = 0.5$ 与 $H_a: \pi \neq 0.5$ 进行假设检验,其中 $y = 0$ 且 $n = 25$。前面我们指出了计分统计量为 $z = -5.0$。利用零假设时的二项分布 bin(25,0.5),相应统计量对应的精确 P 值为

$$P(|z| \geq 5.0) = P(Y = 0 \text{ 或 } Y = 25) = 0.5^{25} + 0.5^{25} = 0.000\,000\,06。$$

$100(1-\alpha)\%$ 的置信区间包括所有在二项分布精确检验中 P 值超过 α 的 π_0。目前最优的区间(Clopper and Pearson, 1934)是根据*尾部法*(*tail method*)所构建的置信区间。该方法要求计算分布的每个尾部超过 $\alpha/2$ 的 P 值。区间的上限和下限可由在以下方程中关于 π_0 的解求得:

$$\sum_{k=y}^{n} \binom{n}{k} \pi_0^k (1-\pi_0)^{n-k} = \alpha/2 \quad \text{以及} \quad \sum_{k=0}^{y} \binom{n}{k} \pi_0^k (1-\pi_0)^{n-k} = \alpha/2,$$

特别地,当 $y = 0$ 时,下限为 0;当 $y = n$ 时,上限为 1。当 $y = 1, 2, \cdots, n-1$ 时,根据二项分布之和与不完全 β 函数(incomplete beta function)之间的联系以及相应的 β 累积分布函数和 F 分布,相应置信区间等于

$$\left[1 + \frac{n-y+1}{yF_{2y,2(n-y+1)}(1-\alpha/2)}\right]^{-1} < \pi < \left[1 + \frac{n-y}{(y+1)F_{2(y+1),2(n-y)}(\alpha/2)}\right]^{-1},$$

其中 $F_{a,b}(c)$ 表示自由度为 a 和 b 的 F 分布的第 $1-c$ 个分位数。当 $y = 0$ 且 $n = 25$ 时,关于 π 的 Clopper-Pearson 置信区间为(0.0,0.137)。

从理论上来看,这一方法近乎完美。然而,它存在一个严重的问题。由于离散性的存在,置信区间对于任何 π 的实际涵盖概率都至少不小于名义上的置信水平(Casella and Berger, 2001: 434; Neyman, 1935),而且可能会比后者大很多。类似地,对于在一个固定的 α 水平(如 0.05)上检验 $H_0: \pi = \pi_0$,有时可能没法达到指定的置信水平。由于可能的样本数量是有限的,因此 P 值的可能取值的数量也是有限的,其中可能不包括 0.05。当需要针对固定的 π_0 去检验 H_0 时,我们可以选取一个可以发生的特定 α 作为 P 值。然而,对于区间估计而言,这种方法并不可行。这是因为构建置信区间意味着将 $H_0: \pi = \pi_0$ 中所有 π_0 的取值范围进行转换,并且每个不同的 π_0 值都可以对应一系列可能的 P 值;也即,在一次检验中可能不存在一个唯一的零假设参数值 π_0。

对于任何给定的参数值,置信区间所涵盖的实际概率有可能远大于名义上的置信水平。图 1.3 给出了当 $n = 25$ 时,由 Clopper-Pearson 法、计分法,以及沃尔德法计算的置信区间所涵盖的概率随 π 变动的情况。在给定 π 值时,任一方法计算的置信区间所涵盖的概率等于相应区间中包含 π 的所有可能样本的二项分布概率之和。这里总共存在 26 个可能的样本以及 26 个相应的置信区间,因此,所涵盖的概率可能对应着 0 到 26 个二项分布概率的求和。当 π 的值从 0 到 1 发生变动时,所涵盖的概率会随着 π 移入或移出某个区间而上升或下降。图 1.3 显示,沃尔德法计算的置信区间涵盖的概率太低,而 Clopper-Pearson 法计算

① 跳过标有*的章节基本不会影响对本书的阅读。

的区间则刚好相反。除非 π 的取值非常接近于 0 或 1，计分法的结果相对较好。相应的置信区间所涵盖的概率一般接近于名义水平，而不是系统性地偏高或偏低。因而，在 π 的值不是非常接近 0 或 1 时，计分法是一种很好的方法(习题1.23)。

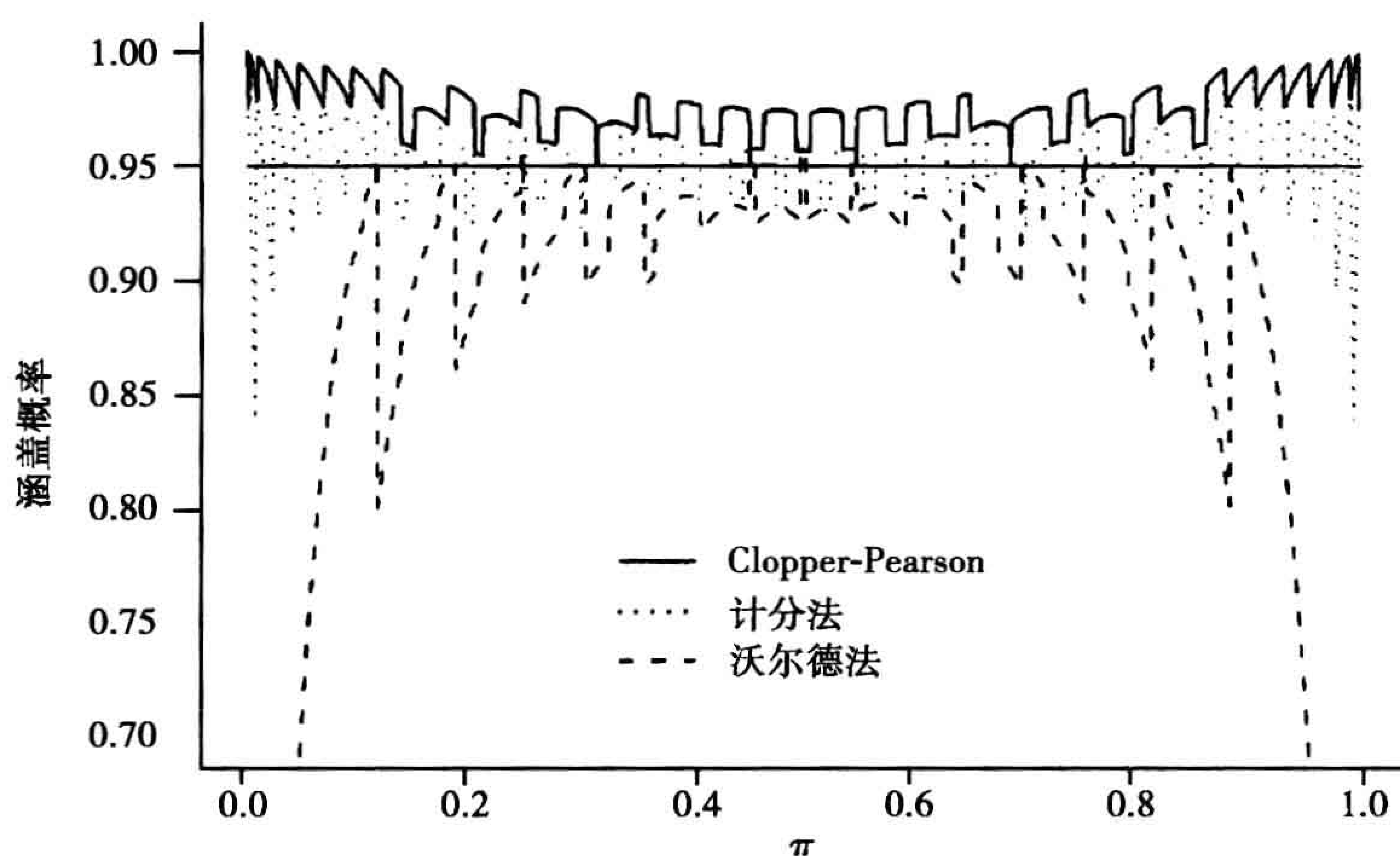


图 1.3 当 $n=25$ 时二项分布参数 π 的 95% 的名义置信区间所涵盖的概率

在使用小样本分布所导致的离散性问题中，通过一个双边检验(而不是两个单边检验)的转换所得到的置信区间往往较短。这时，置信区间就是在双边检验中 P 值大于 α 时的一组参数值。对于二项分布参数而言，有关方法可参见：Blaker(2000)、Blyth and Still(1983)、Sterne(1954)。针对所观察到的结果 y_o，Blaker 方法的 P 值等于两个单尾二项分布概率 $P(Y\geq y_o)$ 和 $P(Y\leq y_o)$ 中较小的一个值加上在另一尾部所能达到的最接近但不大于该单尾概率的概率值。尽管有软件可以用来计算相应的置信区间(Blaker 给出了 S-Plus 的公式)，但这些计算仍相当复杂。运用该方法所得到的结果仍然相对保守，但会比 Clopper-Pearson 区间好一些。拿素食主义者的例子来说，用 Blaker 精确法所计算的 95% 置信区间为(0.0,0.128)，而相应的 Clopper-Pearson 区间为(0.0,0.137)。

1.4.5 基于中位 P 值的统计推断*

为了修正小样本分布中的离散性问题，我们可以利用*中位 P 值*(*mid-P-value*)进行统计推断(Lancaster, 1961)。对于观察值为 t_o 的检验统计量 T，在对 H_a: $T>t_o$ 进行单边检验时，

$$中位 P 值 = \frac{1}{2}P(T=t_o)+P(T>t_o),$$

式中的概率是根据零分布计算而来的。因此，中位 P 值比普通的 P 值小，二者之差为所观察到的结果发生的概率的一半。与普通的 P 值相比，中位 P 值在特性上更像一个具有连续分布的检验统计量的相应 P 值。它的两个单边检验的 P 值之和等于 1.0。尽管是离散的，在 H_0 下它的零分布更像连续情况下的均匀分布。例如，它的零假设期望值是 0.5，而对离散检验统计量的普通 P 值而言，该期望值大于 0.5。

与精确检验中的普通 P 值不同，使用中位 P 值所进行的检验并不保证第一类误差(type I error)的发生概率不超过名义置信水平(习题 1.19)。然而，它通常都比较准确，一般结果会稍偏保守。与普通的精确检验相比，中位 P 值偏保守的程度要低一些。同样地，我们可以通过转换根据中位 P 值的精确分布所进行的检验来构建较不保守的置信区间(比如说，95% 的置信区间是中位 P 值大于 0.05 的一组参数值)。

在 $y=0$ 且 $n=25$ 的素食主义者所占比例的例子中，对 $H_0: \pi=0.5$ 和 $H_a: \pi\neq 0.5$ 进行检验，所观察到的结果是一种最极端的可能情况。因而，中位 P 值等于普通 P 值的一半，即 0.000 000 03。利用有关中位 P 值的二项分布精确检验，通过 Clopper-Pearson 法计算的关于 π 的 95% 置信区间为（0.000，0.113），小于普通 Clopper-Pearson 法的区间（0.000，0.137）。

中位 P 值法是在过于保守的区间估计和通过无关的随机处理来消除离散性问题的方法之间一种合理的折中。在分析高度离散的分布时，我们建议使用这种方法来进行统计检验与构建置信区间。

1.5　多项分布参数的统计推断

我们现在来介绍对多项分布参数 $\{\pi_j\}$ 的统计推断。在 n 个观测值中，n_j 表示第 j 类结果发生的次数，$j=1,\cdots,c$。

1.5.1　多项分布参数的估计

首先，我们给出关于 $\{\pi_j\}$ 的最大似然估计值。作为 $\{\pi_j\}$ 的一个函数，多项分布的概率密度函数（式 1.2）与其核函数

$$\prod_j \pi_j^{n_j} \quad (\text{其中所有的 } \pi_j \geqslant 0 \text{ 且} \sum_j \pi_j = 1) \tag{1.14}$$

成比例。最大似然估计值就是使式 1.14 取最大值的 $\{\pi_j\}$。

多项分布的对数似然函数为

$$L(\pi) = \sum_j n_j \log \pi_j\text{。}$$

由于 $\pi_c = 1-(\pi_1+\cdots+\pi_{c-1})$，去掉冗余项，$L$ 可视为 $(\pi_1,\cdots,\pi_{c-1})$ 的函数。进而，$\partial\pi_c/\partial\pi_j = -1, j=1,\cdots,c-1$。

由于

$$\frac{\partial \log \pi_c}{\partial \pi_j} = \frac{1}{\pi_c}\frac{\partial \pi_c}{\partial \pi_j} = -\frac{1}{\pi_c},$$

对 $L(\pi)$ 关于 π_j 求导，便得似然方程：

$$\frac{\partial L(\pi)}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0\text{。}$$

最大似然估计的解满足 $\hat{\pi}_j/\hat{\pi}_c = n_j/n_c$。这样

$$\sum_j \hat{\pi}_j = 1 = \frac{\hat{\pi}_c(\sum_j n_j)}{n_c} = \frac{\hat{\pi}_c n}{n_c},$$

因此，$\hat{\pi}_c = n_c/n$ 且 $\hat{\pi}_j = n_j/n$。这组解最大化了似然函数，这可以从本书后面（第 8.6 节）介绍的一般性结果得以印证。因而，对 $\{\pi_j\}$ 的最大似然估计就是相应的样本比例。

1.5.2　检验一个特定多项分布的皮尔逊统计量

1900 年，英国著名统计学家卡尔·皮尔逊（Karl Pearson）提出了一种假设检验，这是最早出现的统计推断方法之一。该检验用于描述关联关系，对分类数据的分析产生了革命性的影响。皮尔逊统计检验考察多项分布的参数是否等于一组特定的值。他提出该

检验的初衷是分析一次蒙特卡洛轮盘赌的结果是否具有相等的可能性(Stigler, 1986)。

考虑 $H_0: \pi_j = \pi_{j0}, j = 1, \cdots, c$,其中 $\sum_j \pi_{j0} = 1$。当 H_0 成立时,$\{n_j\}$ 的期望值被称为*期望频数*(*expected frequencies*),即 $\mu_j = n\pi_{j0}, j = 1, \cdots, c$。皮尔逊提出的检验统计量可表示为

$$X^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j}。\tag{1.15}$$

对于固定的 n,$\{n_j - \mu_j\}$ 的差越大,X^2 值也越大。令 X_o^2 表示所观察到的 X^2 值。检验的 P 值就是 $P(X^2 \geqslant X_o^2)$ 的零分布值。这相当于在所有的可能频数结果中(总和为 n),满足 $X^2 \geqslant X_o^2$ 的多项零分布的概率之和。

在大样本的情况下,X^2 近似服从自由度为 $\mathrm{df} = c - 1$ 的卡方分布。P 值可以近似为 $P(\chi_{c-1}^2 \geqslant X_o^2)$,其中 χ_{c-1}^2 代表自由度为 $\mathrm{df} = c - 1$ 的随机卡方变量。式 1.15 所表示的统计量被称作*皮尔逊卡方统计量*(*Pearson chi-squared statistic*)。

1.5.3 例子:对孟德尔理论的检验

在皮尔逊检验的诸多应用中,其中之一是在遗传学中用它来检验孟德尔(Mendel)关于自然遗传的理论。孟德尔将纯黄种的与纯绿种的豌豆交叉耕种。他预测第二代的杂交种子会是75%的黄豌豆和25%的绿豌豆,黄种系占据主导地位。在一次实验中共生产了 $n = 8\ 023$ 个种子,其中黄种子为 $n_1 = 6\ 022$ 个,绿种子为 $n_2 = 2\ 001$ 个。在 $H_0: \pi_{10} = 0.75, \pi_{20} = 0.25$ 下对应的期望频数为 $\mu_1 = 8\ 023(0.75) = 6\ 017.25$ 以及 $\mu_2 = 8\ 023(0.25) = 2\ 005.75$。皮尔逊统计量 $X^2 = 0.015(\mathrm{df} = 1)$ 所对应的 P 值为 $P = 0.90$。该结果与孟德尔的假设并不矛盾。

孟德尔进行了几次这样的实验。在 1936 年,R. A. Fisher 对孟德尔的结果进行了总结。他利用了卡方分布的可复制性特点:如果 $X_1^2, \cdots, X_k^2$ 分别为具有 $v_1, \cdots, v_k$ 个自由度的独立的卡方统计量,那么它们的和 $\sum_i X_i^2$ 也服从一个 $\mathrm{df} = \sum_i v_i$ 的卡方分布。Fisher 得到了一个总和为 42 的卡方统计量,其 $\mathrm{df} = 84$。对于 $\mathrm{df} = 84$ 的卡方分布,它的均值为 84,标准差为 $(2 \times 84)^{1/2} = 13.0$,因而其超过 42 的右尾概率为 $P = 0.999\ 96$。换句话说,卡方统计量是如此之小,以至于检验的拟合看上去似乎太好了。

Fisher 评论道:"孟德尔的期望与他报告的结果之间的这种一致性显示,这比在进行了几千次重复实验中选取的最好的结果还要好……我认定孟德尔肯定是被他的园艺助理欺骗了,这个助理太知道在每一次试验中他的老板想得到怎样的结果。"在当时的一封信中(参见 Box, 1978:297),他提到:"现在,当数据是假造的时候,我非常清楚地知道人们一般总是低估频数的纯粹随机变动,以至于他们总倾向于使假造的数据与期望的结果过于一致。"总之,拟合优度检验不仅能展现一个拟合结果够不够充分,也能够揭示拟合结果比我们在随机波动情况下所预期的还要好的情形[R. A. Fisher 的女儿,John Fisher Box(1978, p. 295-300),以及 Freedman 等(1978, pp. 420-428, 478)讨论了 Fisher 对孟德尔数据的分析以及有关的争论。尽管孟德尔的数据可能存在问题,但是后续的研究结果还是证实了他的理论]。

1.5.4 卡方检验的理论推导*

我们现在概述为什么皮尔逊统计量的极限服从卡方分布。对于规模为 n 的多项分

布样本$(n_1,\cdots,n_c)$而言,n_j 的边际分布是服从 $\text{bin}(n,\pi_j)$ 的二项分布。当 n 很大时,按照对二项分布的正态近似,n_j(以及 $\hat{\pi}_j=n_j/n$)近似服从正态分布。更一般地说,按照中心极限定理,样本比例 $\hat{\pi}=(n_1/n,\cdots,n_{c-1}/n)'$近似服从多元正态分布(第 14.1.4 节)。令 Σ_0 表示$\sqrt{n}\hat{\pi}$ 的零分布协方差矩阵,并令 $\pi_0=(\pi_{10},\cdots,\pi_{c-1,0})'$。在 H_0 下,由于 $\sqrt{n}(\hat{\pi}-\pi_0)$ 收敛于 $N(0,\Sigma_0)$分布,则二次项

$$n(\hat{\pi}-\pi_0)'\Sigma_0^{-1}(\hat{\pi}-\pi_0) \tag{1.16}$$

的分布收敛于自由度为 $\text{df}=c-1$ 的卡方分布。

在第 14.1.4 节我们会讲到,$\sqrt{n}\hat{\pi}$ 的协方差矩阵具有如下元素:

$$\sigma_{jk}=\begin{cases}-\pi_j\pi_k & (j\neq k)\\ \pi_j(1-\pi_j) & (j=k)\end{cases}。$$

当$j\neq k$ 时,矩阵 Σ_0^{-1} 的第(j,k)个元素为 $1/\pi_{c0}$;当 $j=k$ 时,相应元素为$(1/\pi_{j0}+1/\pi_{c0})$(读者可以通过证明 $\Sigma_0\Sigma_0^{-1}$ 等于单位矩阵来对此进行验证)。代入上述结果,计算(合并同类项)可得,公式 1.16 简化为 X^2。更为一般情况下的正式证明,将留在本书第 14.3 节介绍。

上述推导与皮尔逊在 1900 年的论断很相似。R. A. Fisher(1922)给出了一个更简单的推导,其精髓如下:假定$(n_1,\cdots,n_c)$是均值为$(\mu_1,\cdots,\mu_c)$的独立的泊松随机变量。在$\{\mu_j\}$较大的情况下,其标准化值$\{z_j=(n_j-\mu_j)/\sqrt{\mu_j}\}$近似服从标准正态分布。因而,$\sum_j z_j^2=X^2$ 近似服从自由度为c的卡方分布。加上一个线性限定条件 $\sum_j(n_j-\mu_j)=0$,泊松分布转化成多项分布,从而损失了一个自由度。

当 $c=2$ 时,皮尔逊 X^2 简化为标准计分统计量(式 1.11)的平方。对于孟德尔的数据来说,$\hat{\pi}_1=6\ 022/8\ 023$,$\pi_{10}=0.75$,$n=8\ 023$,以及 $z_S=0.123$,相应的 $X^2=(0.123)^2=0.015$。在一般情况下,皮尔逊检验实际上就是关于多项分布参数的计分检验。

1.5.5 似然比卡方

另一种关于多项分布参数的检验是似然比检验。多项分布似然函数的核函数为公式 1.14。在 H_0 下,似然函数在 $\hat{\pi}_j=\pi_{j0}$时达到最大值。在一般的情况下,当 $\hat{\pi}_j=n_j/n$ 时,似然函数实现最大值。两个似然函数之比等于

$$\Lambda=\frac{\prod_j(\pi_{j0})^{n_j}}{\prod_j(n_j/n)^{n_j}}。$$

因此,由 G^2 所表示的似然比统计量等于

$$G^2=-2\log\Lambda=2\sum n_j\log(n_j/n\pi_{j0})。\tag{1.17}$$

这一统计量具有式 1.12 的形式,被称为*似然比卡方统计量*(*likelihood-ratio chi-squared statistic*)。G^2 的值越大,拒绝 H_0 的可能性也越大。

一般情况下,参数空间包括了$\{\pi_j\}$且满足 $\sum_j\pi_j=1$,因此参数空间的维度是 $c-1$。在 H_0 下,$\{\pi_j\}$是完全给定的,因而空间的维度为 0。两个参数空间维度之差等于$(c-1)$。在大样本的情况下,G^2 服从 $\text{df}=c-1$ 的卡方零分布。

当 H_0 成立时,皮尔逊 X^2 和似然比 G^2 都渐近服从 $\text{df}=c-1$ 的卡方分布。事实上,二

者在这种情况下是渐近等价的；具体来说，X^2-G^2 依概率收敛于零（详见第 14.3.4 节）。当 H_0 不成立时，二者之差一般随着 n 成比例增加；即使在 n 非常大的情况下，二者也可能存在很大差别。

在 c 固定的情况下，随着 n 的增加，通常 X^2 会比 G^2 更快地收敛于卡方分布。当 $n/c<5$ 时，G^2 与卡方分布的相似度一般较差。当 c 很大时，如果各类别间期望频数的分布相对均匀，即便 n/c 的值小到 1 时，X^2 与卡方分布的相似性都可以接受。我们将在第 9.8.4 节对此进行更详细的讨论。另外，也可以根据多项分布的概率公式计算这些检验统计量的精确分布（Good et al.，1970）。

1.5.6 利用估计的期望频数进行检验

皮尔逊 X^2（式 1.15）将一个样本分布与假设的分布 $\{\pi_{j0}\}$ 进行比较。在某些应用中，$\{\pi_{j0}=\pi_{j0}(\boldsymbol{\theta})\}$ 是一组未知参数 $\boldsymbol{\theta}$ 的函数。关于 $\boldsymbol{\theta}$ 的最大似然估计值 $\hat{\boldsymbol{\theta}}$ 决定了对 $\{\pi_{j0}\}$ 的最大似然估计值 $\{\pi_{j0}(\hat{\boldsymbol{\theta}})\}$，进而也决定了在 X^2 中对期望频数的最大似然估计值 $\{\hat{\mu}_j=n\pi_{j0}(\hat{\boldsymbol{\theta}})\}$。这样，用估计值 $\{\hat{\mu}_j\}$ 代替 $\{\mu_j\}$ 会影响 X^2 的分布。当 $\dim(\boldsymbol{\theta})=p$ 时，正确的自由度为 $df=(c-1)-p$（详见第 14.3.3 节）。皮尔逊本人并没有意识到这一问题（第16.2 节）。

现在我们介绍利用估计的期望频数来进行拟合优度检验。例如，出生在佛罗里达州奥基巧比郡（Okeechobee County，Florida）牧场的 156 只小牛，按照是否在出生 60 天内得过肺部感染进行了分组。那些得过肺部感染的小牛又按照是否在第一次感染痊愈的两周内又发生了第二次感染进行了划分。表 1.1 显示了这些数据。没有得过初次感染的小牛当然也不会得到二次感染，因此初次感染为“否”而二次感染为“是”的频数为零。这种情况被称作*结构性零值*（*structural zero*）。

表 1.1 小牛发生初次与二次肺部感染的情况

初次感染	二次感染[a]	
	是	否
是	30(38.1)	63(39.0)
否	0(—)	63(78.9)

来源：数据由佛罗里达大学兽医学院 Thang Tran 和 G. A. Donovan 友情提供。

a 括号中的值为估计的期望频数。

本项研究的一个目的是检验小牛发生初次感染的概率是否与初次感染后又发生二次感染的条件概率相同。换句话说，如果由 π_{ab} 表示一只小牛被划分为表中第 a 行、第 b 列的概率，那么零假设为

$$H_0: \pi_{11}+\pi_{12}=\pi_{11}/(\pi_{11}+\pi_{12})$$

或者 $\pi_{11}=(\pi_{11}+\pi_{12})^2$。令 $\pi=\pi_{11}+\pi_{12}$ 表示发生初次感染的概率。零假设是说这些概率满足表 1.2 所显示的结构，即，关于初次感染/二次感染的类别（是—是，是—否，否—否）的三项分布的概率等于（$\pi^2,\pi(1-\pi),1-\pi$）。

令 n_{ab} 表示落在类别（a,b）内的观测值数量。对 π 的最大似然估计值就是使多项分布似然函数的核函数

$$(\pi^2)^{n_{11}}(\pi-\pi^2)^{n_{12}}(1-\pi)^{n_{22}}。$$

最大化时 π 的取值。

表 1.2 假设的概率结构

初次感染	二次感染		
	是	否	小 计
是	π^2	$\pi(1-\pi)$	π
否	—	$1-\pi$	$1-\pi$

相应的对数似然函数为

$$L(\pi) = n_{11}\log \pi^2 + n_{12}\log(\pi - \pi^2) + n_{22}\log(1-\pi)。$$

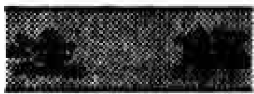
对 π 求导,便得似然方程

$$\frac{2n_{11}}{\pi} + \frac{n_{12}}{\pi} - \frac{n_{12}}{1-\pi} - \frac{n_{22}}{1-\pi} = 0。$$

其解为

$$\hat{\pi} = (2n_{11} + n_{12})/(2n_{11} + 2n_{12} + n_{22})。$$

由表 1.1 可知,$\hat{\pi}=0.494$。由于 $n=156$,所估计的期望频数分别为 $\hat{\mu}_{11}=n\hat{\pi}^2=38.1$,$\hat{\mu}_{12}=n(\hat{\pi}-\hat{\pi}^2)=39.0$,$\hat{\mu}_{22}=n(1-\hat{\pi})=78.9$。表 1.1 给出了这些值。皮尔逊统计量为 $X^2=19.7$。由于有 $c=3$ 种可能出现的结果,并且有 $p=1$ 个参数(π)决定了期望频数,自由度为 $\mathrm{df}=(3-1)-1=1$。该数据强烈拒绝了 H_0($P=0.000\ 01$)。对表 1.1 的分析显示,发生初次感染的小牛要比 H_0 所预测的多得多,而发生二次感染的则少得多。该项研究的结论是,初次感染具有一种免疫效应,因而会降低发生二次感染的可能性。



第 1.1 节:分类数据

1.1 Stevens(1951)给出了关于测量尺度(定类、定序、定距)的定义。其他的测量尺度来自于以上尺度的混合。例如,*部分定序*(*partially ordered*)尺度是指当研究对象的回答除了“不知道”或“还未决定”等类别外,其他的类别存在排序的情况。

第 1.3 节:分类数据的统计推断

1.2 计分法并不使用 $\hat{\beta}$。因而,当 $\hat{\beta}$ 是模型参数时,我们通常不需要拟合模型就可以计算计分统计量来检验 $H_0:\beta=\beta_0$。当在探索性分析中需要拟合多个模型并且模型拟合需要大量运算时,这是计分法很重要的优点。有关计分法和似然比方法的一个优点是,即便在 $|\hat{\beta}|=\infty$ 时,它们仍然成立。而在这种情况下,我们无法计算沃尔德统计量。沃尔德方法的另一个缺点是它的结果取决于参数化的选择,且基于 $\hat{\beta}$ 及其标准误的统计推断与基于其非线性函数如 $\log\hat{\beta}$ 及其标准误的统计推断并不等价。

第 1.4 节:二项分布参数的统计推断

1.3 与其他人一起,Agresti 和 Coull(1998)、Blyth 和 Still(1983)、Brown 等(2001)、Ghosh(1979),以及 Newcombe(1998a)分析了关于参数 π 的计分区间与沃尔德区间相比的优越性。在精确方法中,Blaker 的方法(2000)具有非常好的特性。它落在 Clopper-Pearson 区间的里面,并且具有嵌套的特性,即名义水平较高的置信区间总会包含所有较低水平的区间。

1.4 利用对大样本方法的连续性修正可以作为对精确小样本方法的近似。因而,有关

的结果一般会相对保守。我们在这里不介绍这些方法,原因是如果你偏好一种精确方法,基于现有的运算能力,你可以直接使用该方法而不需要去选择某种对它的近似。

1.5 理论上,我们可以通过对临界区域的边界进行辅助性随机化分配来消除检验中的离散性问题(参见习题 1.19)。为了在给定的概率水平下拒绝在边界上的零假设,即使在某个 P 值无法实现的情况下,我们可以得到一个给定的总的第一类误差(type I error)水平 α。在这样的随机法中,单边 P 值为

$$\text{随机化}\,P\,\text{值} = U \times P(T = t_0) + P(T > t_0),$$

其中 U 表示服从(0,1)均匀分布的随机变量(Stevens, 1950)。在应用中,让一个随机数来影响决定很荒谬,因而这种方法价值不大。中位 P 值法用其期望值代替了这种随意的均匀乘数 $U \times P(T = t_0)$。

第 1.5 节:多项分布参数的统计推断

1.6 卡方分布的均值为 df,方差为 2df,并且斜度等于 $(8/\text{df})^{1/2}$。当 df 很大时,它近似于正态分布。Greenwood 和 Nikulin(1996)、Kendall 和 Stuart(1979),以及 Lancaster(1969)介绍了关于卡方分布的其他特性。Cochran(1952)对卡方拟合检验的发展进行了历史回顾。另见:Cressie and Read(1989)、Koch and Bhapkar(1982)、Koehler(1998)、Moore(1986b)。

习　题

应用部分

1.1 指出下列每个变量分别属于定类、定序或定距中的哪一种?

a. 英国人的政党倾向(工党、保守党、社会民主党)

b. 焦虑程度(无、轻度、中度、严重、非常严重)

c. 患者的生存时间(以月为单位)

d. 诊所位置(伦敦、波士顿、麦迪逊、罗彻斯特、蒙特利尔)

e. 化疗治疗肿瘤的效果(完全消失、部分减小、没有变化、继续增大)

f. 喜欢的饮品(水、果汁、牛奶、碳酸饮料、啤酒、红酒)

g. 公司存货水平的评估(太低、正常、太高)

1.2 一次考试中共包括 100 道选择题。每题有四个可选答案,其中一个是正确的。对于每个问题,一个学生通过随机猜测选择答案。

a. 给出该学生回答正确的题目数量的分布。

b. 求出该分布的均值和方差。如果该学生答对了至少 50 道题目,这令人惊讶吗?为什么?

c. 给出 (n_1, n_2, n_3, n_4) 的分布,其中 n_j 表示学生选择第 j 个答案的次数。

d. 求 $E(n_j)$, $\text{var}(n_j)$, $\text{cov}(n_j, n_k)$,以及 $\text{corr}(n_j, n_k)$。

1.3 一项实验分别研究几组总量为 n 的昆虫在使用一定量的杀虫剂之后能存活下来的数量。在实验中昆虫的存活会受到与组别相关的其他未测量的因素(如温度)的影响。说明为什么在实验中每组昆虫存活数量的分布与 $\text{bin}(n, \pi)$ 分布相比可能存在过度离散。

1.4 英国作家 Graham Greene 在他的自传《一种人生》中描述了一段他玩俄罗斯轮盘赌的经历。该"游戏"将一颗子弹放入一把手枪的六个枪膛之一,旋转枪膛并随机选

取一个,然后冲着某人的头部射击。

a. Greene 共玩了六次这个游戏,非常幸运的是没有一次射出子弹。求这一结果发生的概率。

b. 假设他继续进行游戏直到子弹射出为止。令 Y 表示当子弹射出时玩该游戏的次数。给出 Y 的概率密度函数并加以说明。

1.5 考虑以下这句话:"请告诉我您认为是否应该允许一个怀孕的妇女在已婚并且不想再要更多孩子时进行合法的流产。"在美国民意研究中心(National Opinion Research Center,NORC)1996 年进行的综合社会调查(General Social Survey)中,842 名被访者选择了"是",并有 982 名选择了"否"。令 π 表示总体中选择"是"的比例。运用计分法求对 $H_0: \pi=0.5$ 进行检验的 P 值,并构建关于 π 的 95% 置信区间。对结果加以解释。

1.6 参考第 1.4.3 节中素食主义者的例子。对于检验备择假设为 $H_a: \pi \neq 0.5$ 的假设 $H_0: \pi=0.5$,证明:

a. 似然比统计量等于 $2[25 \log(25/12.5)]=34.7$。

b. 卡方形式的计分统计量等于 25.0。

c. 沃尔德统计量 z 或其卡方形式等于无穷大。

1.7 在一项关于比较新药品与标准药品的交叉试验中,π 表示结果显示新药品效果更好的概率。我们想要估计 π 并检验假设 $H_0: \pi=0.5$ 及备择假设 $H_a: \pi \neq 0.5$。在 20 次相互独立的观测中,每次都是新药品的效果更好。

a. 给出似然函数,并画图说明。给出关于 π 的最大似然估计值。

b. 对 π 进行沃尔德检验,并构建 95% 的沃尔德置信区间。这些结果有意义吗?

c. 进行计分检验,报告 P 值。构建 95% 的计分置信区间。对结果加以解释。

d. 进行似然比检验,并构建相应的 95% 置信区间。对结果加以解释。

e. 进行二项分布精确检验,并给出相应的 95% 置信区间。对结果加以解释。

f. 假设研究者想要一个足够大的样本以保证在置信水平 0.95 上所估计的偏好新药品的概率的变动不超过 0.05。如果概率的真值为 0.90,大概需要多大的样本?

1.8 一项关于玉米的叶绿素遗传的实验结果为,在 1 103 株自受精的杂交玉米中,854 株是绿色的,249 株是黄色的。理论所预测的绿色与黄色的比是 3:1。对 3:1 是真实的比率进行假设检验。报告检验的 P 值,并加以解释。

1.9 表 1.3 所给出的是 Ladislaus von Bortkiewicz 关于普鲁士军队中被军中骡子踢死的士兵死亡数据(Fisher, 1934; Quine and Seneta, 1987)。该数据对 10 个军营进行了长达 20 年的观测。在 109 个军营—年的观测中没有发生死亡,在 65 个军营—年中发生了 1 例死亡,等等。对均值进行估计,并检验是否在这五个类别中事件发生的概率服从泊松分布(将 ≥4 的结果进行合并)。

表 1.3 习题 1.9 的数据

死亡数量	军营—年数
0	109
1	65
2	22
3	3
4	1
≥5	0

1.10 一个样本包括100名受到痛经困扰的妇女。有种新的止痛剂宣称其效果要好于标准的止痛剂。在一项分别使用不同止痛剂的交叉实验中,40人报告标准止痛剂的效果更好,另外60人报告新止痛剂的效果更好。对这一数据加以分析。

理论与方法

1.11 为什么当二项分布参数 π 的值接近于0或1时,比取值接近于$\frac{1}{2}$时更容易得到一个精确的估计?

1.12 假定 $P(Y_i=1)=1-P(Y_i=0)=\pi, i=1,\cdots,n$,其中$\{Y_i\}$相互独立。令 $Y=\sum_i Y_i$。

a. $\text{var}(Y)$和 Y 的分布分别是什么?

b. 当$\{Y_i\}$不独立而是存在着 $\rho>0$ 的两两相关时,证明 $\text{var}(Y)>n\pi(1-\pi)$,即与二项分布相比存在过度离散(Altham(1978)讨论了允许相关试验情况下二项分布的一般情形)。

c. 假设存在异质性:即对所有 i 来说,$P(Y_i=1|\pi)=\pi$ 是一个随机变量,其密度函数 $g(\cdot)$ 的取值范围为[0,1],均值为 ρ,方差为正。证明 $\text{var}(Y)>n\rho(1-\rho)$(当 π 服从β分布(beta distribution)时,Y 服从第13.3节中介绍的β-*二项分布*(*beta-binomial distribution*))。

d. 假定 $P(Y_i=1|\pi_i)=\pi_i, i=1,\cdots,n$,其中$\{\pi_i\}$独立于 $g(\cdot)$。解释为什么 Y 无条件地而不是在给定$\{\pi_i\}$的条件下服从 $\text{bin}(n,\rho)$ 分布(提示:在每一种情况下,Y 等于一系列独立同分布的伯努利试验(Bernoulli trails)之和吗?)。

1.13 对于一系列相互独立的伯努利试验,Y 是在第 k 次失败发生前的成功次数。解释为什么它的概率密度函数是*负二项分布*(*negative binomial*),

$$p(y)=\frac{(y+k-1)!}{y!(k-1)!}\pi^y(1-\pi)^k,\quad y=0,1,2,\cdots。$$

(对此函数,$E(Y)=k\pi/(1-\pi)$,$\text{var}(Y)=k\pi/(1-\pi)^2$,因此 $\text{var}(Y)>E(Y)$;泊松分布是它在 $k\pi=\mu$ 固定不变的情况下,当 $k\to\infty$ 且 $\pi\to 0$ 时的极限分布。)

1.14 对于多项分布,证明

$$\text{corr}(n_j,n_k)=-\pi_j\pi_k/\sqrt{\pi_j(1-\pi_j)\pi_k(1-\pi_k)}。$$

证明当 $c=2$ 时,$\text{corr}(n_1,n_2)=-1$。

1.15 证明二项分布的矩量生成函数(moment generating function, mgf)为 $m(t)=(1-\pi+\pi e^t)^n$,并利用该函数求出前两个矩量。证明泊松分布的矩量生成函数是 $m(t)=\exp\{\mu[\exp(t)-1]\}$,并用其求出前两个矩量。

1.16 令似然比统计量为 t_o。在最大似然估计值处,证明在假设 H_a 下数据发生的可能性是在假设 H_0 下的 $\exp(t_o/2)$倍。

1.17 假定 $y_1,y_2,\cdots,y_n$ 服从一个相互独立的泊松分布。

a. 给出它的似然函数。证明最大似然估计值 $\hat{\mu}=\bar{y}$。

b. 使用以下方法构建关于 $H_0:\mu=\mu_0$的大样本检验统计量:(i)沃尔德法,(ii)计分法,(iii)似然比法。

c. 使用以下方法构建关于 μ 的大样本置信区间:(i)沃尔德法,(ii)计分法,(iii)似然比法。

1.18 关于泊松分布参数的统计推断常常基于它与二项分布以及多项分布的联系。证

明如何通过对二项分布参数 π 的相应检验,来对两个服从泊松分布(y_1,y_2)的独立总体进行检验 $H_0:\mu_1=\mu_2$(提示:在 $n=y_1+y_2$ 的条件下,确定 $\pi=\mu_1/(\mu_1+\mu_2)$)。如何根据 π 的置信区间构建关于 μ_1/μ_2 的置信区间?

1.19 通常,研究者使用名义置信水平为 P(第一类错误)$=0.05$ 进行检验,即当 P 值 $\leqslant 0.05$ 时拒绝 H_0。使用检验统计量 T 进行的一个精确检验具有以下零分布:$P(T=0)=0.30, P(T=1)=0.62, P(T=2)=0.08$,其中 T 取值较大时表示有更大的证据拒绝 H_0。

a. 根据普通的 P 值,指出实际的 P(第一类错误)$=0$。

b. 利用中位 P 值,指出实际的 P(第一类错误)$=0.08$。

c. 当 $P(T=0)=0.30, P(T=1)=0.66, P(T=2)=0.04$ 时,分别求出(a)和(b)中的 P(第一类错误)的值。注意利用中位 P 值的检验既可能过保守,又可能过宽松。使用普通 P 值的精确检验不可能过宽松。

d. 在(a)部分中,随机决策检验(randomized-decision test)生成了一个均匀分布变量 U,取值范围为 $[0,1]$,并且在 $T=2$ 以及 $U\leqslant\frac{5}{8}$ 时拒绝 H_0。证明实际的 P(第一类错误)$=0.05$。这个检验合理吗?

1.20 对于二项分布参数 π,给出如何通过转换来构建沃尔德检验以及计分检验的置信区间。

1.21 在掷硬币时,令 π 表示正面朝上的概率。一项实验通过进行 $n=5$ 次独立的投掷来检验 $H_0:\pi=0.5$ 及备择假设 $H_a:\pi\neq 0.5$。

a. 证明利用二项分布精确检验在 0.05 的置信水平上拒绝 H_0 的零分布真实概率为 0.0,利用大样本计分检验拒绝 H_0 的零分布真实概率为 $\frac{1}{16}$。

b. 假定真实的 $\pi=0.5$。解释为什么 95% 的 Clopper-Pearson 置信区间包含 π 的概率为 1.0(提示:是否存在 y 值使得关于 $H_0:\pi=0.5$ 的两个单边检验都具有 $\leqslant 0.025$ 的 P 值吗?)。

1.22 考虑关于二项分布参数 π 的沃尔德置信区间。由于它在 $\hat{\pi}=0$ 或 1 时不成立,论证当 $0<\pi<1$ 时,该区间包含 π 的概率不可能超过 $[1-\pi^n-(1-\pi)^n]$;因此,当 $0<\pi<1$ 时,区间所涵盖的概率的下确界(infimum)为 0,其大小与 n 无关。

1.23 考虑关于二项分布参数 π 的 95% 的计分置信区间。当 $y=1$ 时,证明该区间的下限大约等于 $0.18/n$;事实上,仅当 $y=0$ 时,π 会落入区间 $0<\pi<0.18/n$。论证在大样本情况下当 π 稍小于 $0.18/n$ 或稍大于 $1-0.18/n$ 时,区间实际涵盖的概率大约为 $e^{-0.18}=0.84$。因而,即便 $n\to\infty$,这一方法也无法保证所涵盖的概率 $\geqslant 0.95$(Agresti and Coull, 1998; Blyth and Still, 1983)。

1.24 根据第 1.4.2 节,关于 π 的计分置信区间的中点 $\tilde{\pi}$ 等于将数据中的每类样本增加 $z_{\alpha/2}^2/2$ 个观测值进行调整后的样本比例。这一思路激发了一种对沃尔德区间的调整,

$$\tilde{\pi}\pm z_{\alpha/2}\sqrt{\tilde{\pi}(1-\tilde{\pi})/n^*},\qquad 其中\ n^*=n+z_{\alpha/2}^2。$$

证明加权平均后的方差 $\tilde{\pi}(1-\tilde{\pi})/n^*$ 不小于在计分区间中根号下的项的方差的加权平均(提示:使用詹生不等式(Jensen's Inequality))。那么,这个区间包含了

计分区间(Agresti 和 Coull(1998)以及 Brown 等(2001)表明,调整后的区间比沃尔德区间更精确,而且它不存在计分区间在接近于 0 或 1 时不准确的缺点(习题 1.23))。

1.25 一个样本规模为 n 的二项分布中成功的次数为 $y=0$ 次。

a. 证明关于 π 的基于似然函数的置信区间为 $[0.0, 1-\exp(z_{\alpha/2}^2/2n)]$。当 $\alpha=0.05$ 时,通过指数函数的展开式证明该区间近似等于 $[0, 2/n]$。

b. 对于计分法,证明其置信区间为 $[0, z_{\alpha/2}^2/(n+z_{\alpha/2}^2)]$,或在 $\alpha=0.05$ 时近似为 $[0, 4/(n+4)]$。

c. 对于 Clopper-Pearson 法,证明置信区间的上限为 $1-(\alpha/2)^{1/n}$,或在 $\alpha=0.05$ 时近似于 $-\log(0.025)/n=3.69/n$。

d. 对于通过中位 P 值调整过的 Clopper-Pearson 法,证明置信区间的上限为 $1-\alpha^{1/n}$,或在 $\alpha=0.05$ 时近似于 $-\log(0.05)/n=3/n$。

1.26 对几何分布(geometric distribution) $p(y)=\pi^y(1-\pi)$, $y=0,1,2,\cdots$,证明尾部法所构建的置信区间(即令 $P(Y\geqslant y)$ 和 $P(Y\leqslant y)$ 等于 $\alpha/2$)为 $[(\alpha/2)^{1/y}, (1-\alpha/2)^{1/(y+1)}]$。证明所有在 0 和 $1-\alpha/2$ 之间的 π 值绝不会落在一个置信区间上,因此该区间实际涵盖的概率超过了 $1-\alpha/2$。

1.27 统计量 T 服从累积分布函数(cdf)为 $F(t)$ 的离散分布。证明 $F(T)$ 在 $[0,1]$ 之间*随机大于*(*stochastically larger*)均匀分布;也即,它的 cdf 在任何一点上都不超过均匀分布的 cdf(Casella and Berger, 2001, p. 77, 434)。说明为什么由此可以推出基于 T 的 P 值的零分布随机大于均匀分布。

1.28 假定 $P(T=t_j)=\pi_j$, $j=1,\cdots$。证明 E(中位 P 值)$=0.5$(提示:证明 $\sum_j \pi_j(\pi_j/2+\pi_{j+1}+\cdots)=(\sum_j \pi_j)^2/2$)。

1.29 统计量 T 具有累积分布函数为 $F(t)$, $p(t)=P(T=t)$,*中位分布函数*(*the mid-distribution function*)为 $F_{\text{mid}}(t)=F(t)-0.5p(t)$(Parzen, 1997)。给定 $T=t_o$,证明中位 P 值等于 $1-F(t_o)$(它也满足 $E[F_{\text{mid}}(T)]=0.5$ 以及 $\text{var}[F_{\text{mid}}(T)]=(1/12)\{1-E[p^2(T)]\}$)。

1.30 基因型 AA, Aa,以及 aa 发生的概率分别为 $[\theta^2, 2\theta(1-\theta), (1-\theta)^2]$。在一个样本规模为 n 的多项分布中,这三种基因型的频数为 (n_1, n_2, n_3)。

a. 构建对数似然函数。证明 $\hat{\theta}=(2n_1+n_2)/(2n_1+2n_2+2n_3)$。

b. 证明 $-\partial^2 L(\theta)/\partial\theta^2=[(2n_1+n_2)/\theta^2]+[(n_2+2n_3)/(1-\theta)^2]$,并且它的期望值是 $2n/\theta(1-\theta)$。利用这一结果推导关于 $\hat{\theta}$ 的渐近标准误。

c. 说明如何检验概率是否真的满足这一模式。

1.31 参考第 1.5.6 节。利用似然函数推出信息矩阵,求 $\hat{\pi}$ 的近似标准误。

1.32 参考第 1.5.6 节。令 a 表示发生了初次、二次和三次感染的小牛数量,b 表示发生了初次、二次但没发生第三次感染的小牛数量,c 表示发生了初次但没发生第二次感染的小牛数量,d 表示没有发生初次感染的小牛数量。令 π 表示小牛发生初次感染的概率。考虑以下假设:给定在时点 $1,\cdots,t-1$ 的感染情况,在时点 t 发生感染的概率也是 π,其中 $t=2,3$。证明 $\hat{\pi}=(3a+2b+c)/(3a+3b+2c+d)$。

1.33 参考公式 1.16 中的二次项。

a. 验证书中所给出的矩阵 Σ_0^{-1} 是 Σ_0 的逆矩阵。

b. 证明公式 1.16 可简化为皮尔逊统计量(式 1.15)。

c. 对于统计量 z_S(式 1.11),证明当 $c=2$ 时,$z_S^2=X^2$。

1.34 在检验 H_0: $\pi_j=\pi_{j0}$,$j=1,\cdots,c$ 时,使用多项分布的样本比例$\{\hat{\pi}_j\}$的似然比统计量(式 1.17)为

$$G^2=-2n\sum_j\hat{\pi}_j\log(\pi_{j0}/\hat{\pi}_j)。$$

证明 $G^2\geqslant 0$,其中当且仅当对所有的 j 都有 $\hat{\pi}_j=\pi_{j0}$ 时,等式才成立(提示：对 $E(-2n\log X)$)应用詹生不等式(Jensen's inequality),其中 X 等于 $\pi_{j0}/\hat{\pi}_j$ 的概率为 $\hat{\pi}_j$)。

1.35 自由度为 $\mathrm{df}=v$ 的卡方分布的矩量生成函数(mgf)为 $m(t)=(1-2t)^{-v/2}$,其中 $|t|<\frac{1}{2}$。通过此函数证明,卡方分布具有可复制特性。

1.36 对于 $c>2$ 的多项分布$(n,\{\pi_j\})$,π_j 的置信极限等于以下方程的解：

$$(\hat{\pi}_j-\pi_j)^2=(z_{\alpha/2c})^2\pi_j(1-\pi_j)/n,\quad j=1,\cdots,c。$$

a. 应用邦弗罗尼不等式(Bonferroni inequality),论证(在大样本的情况下)所有这 c 个区间同时包含$\{\pi_j\}$的概率至少为 $1-\alpha$。

b. 证明 $\hat{\pi}_j-\hat{\pi}_k$ 的标准差等于$[\pi_j+\pi_k-(\pi_j-\pi_k)^2]/n$。在大样本的情况下,说明为什么沃尔德置信区间

$$(\hat{\pi}_j-\hat{\pi}_k)\pm z_{\alpha/2a}\{[\hat{\pi}_j+\hat{\pi}_k-(\hat{\pi}_j-\hat{\pi}_k)^2]/n\}^{1/2}$$

同时包含共 $a=c(c-1)/2$ 个$\{\pi_j-\pi_k\}$的差的概率至少为 $1-\alpha$(参见 Fitzpatrick and Scott, 1987; Goodman, 1965)。

2 对列联表的描述

在本章中,我们介绍描述分类变量之间关系的表格,并给出度量相应关联的参数。第2.2节中介绍的参数可用于比较结果变量各类别在不同组别之间的分布比例。*发生比之比*(*odds ratio*)具有特殊重要的意义,它会在本书后面所讨论的模型中作为重要的参数出现。在第2.3节,我们通过引入第三个变量作为控制变量扩展分析的范围。在引入控制变量后,变量间的关联可能发生巨大的变化。本章主要关注二分变量的情况,即变量只包含两个类别,但是在第2.4节我们将提及描述多类别的定类和定序变量的参数。首先,第2.1节介绍基本的术语和符号。

2.1 列联表的概率结构

两个分类变量的联合分布决定了它们之间的关系,也决定了相应变量的边际分布和条件分布。

2.1.1 列联表及其分布

令 X 和 Y 代表两个分类变量,其中 X 共包括 I 个类别,Y 共包括 J 个类别。将研究对象按照两个变量进行交叉划分则存在 IJ 种可能组合。某一总体中随机选取的研究对象对(X,Y)的回答服从一定的概率分布。这个分布可以通过 I 行和 J 列的矩形表格来表示,其中行表示 X 的类别,列表示 Y 的类别。该表格中的*单元格*(*cell*)代表了 IJ 种可能的结果。当单元格包含的是样本的频数计数时,这样的表格被称为*列联表*(*contingency table*),该术语是由卡尔·皮尔逊(Pearson, 1904)引入的。列联表也叫作*交叉列联表*(*cross-classified table*)。包括 I 行、J 列的列联表被称为 $I\times J$(或 I 乘 J)表格。

表2.1是一个2×3列联表,来自于哈佛医学院的医师健康研究小组发布的一项关于服用阿司匹林与心脏病关系的报告。该研究通过一个5年的随机实验,旨在了解是否常规服用阿司匹林会降低心血管疾病的死亡率。每隔一天,参与此研究的医师服用一片阿司匹林或者安慰剂(placebo),并且研究对象并不知道他们服用的是阿司匹林还是安慰剂。在11 034名服用安慰剂的医师中,18人在研究期间发作了致命性心脏病,然而在11 037名服用阿司匹林的医师中,5人发作了致命性心脏病。

令 π_{ij} 表示(X,Y)落在第 i 行和第 j 列所对应的单元格中的概率。$\{\pi_{ij}\}$ 的概率分布就是 X 和 Y 的*联合分布*(*joint distribution*)。*边际分布*(*marginal distribution*)是对联合分布概率加总所得到的行总计和列总计。我们用 $\{\pi_{i+}\}$ 表示行变量的边际分布,用 $\{\pi_{+j}\}$ 表

示列变量的边际分布,其中下标"+"表示对所指代的变量的求和,即有:

$$\pi_{i+} = \sum_j \pi_{ij} \quad 以及 \quad \pi_{+j} = \sum_i \pi_{ij}。$$

它们满足 $\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1.0$。边际分布给出了每个单变量的分布情况。

表 2.1　关于服用阿司匹林与心肌梗塞的交叉列联表

	心肌梗塞		
	致命性心脏病	非致命性心脏病	未发作
安慰剂	18	171	10 845
阿司匹林	5	99	10 933

来源:Preliminary report:Findings from the aspirin component of the ongoing Physicians' Health Study. *New Engl. J. Med.* 318: 262-264(1988)。

在多数列联表中(如表 2.1),其中一个变量,比如 Y,被视为结果变量,另一个变量(X)被视为解释变量。当 X 给定而不是随机的时候,讨论 X 和 Y 的联合分布没有意义。然而,对于 X 的一个给定类别,Y 服从一个概率分布。这种情况下,可以分析 Y 的分布如何随着 X 的不同取值而变动。给定研究对象关于 X 的取值落在第 i 行,则 $\pi_{j|i}$ 代表 Y 的取值落在第 j 列的概率,$j=1,\cdots,J$。注意,$\sum_j \pi_{j|i} = 1$。概率 $\{\pi_{1|i},\cdots,\pi_{J|i}\}$ 给出了在 X 取第 i 个类别时 Y 的*条件分布*(*conditional distribution*)。比较 Y 在解释变量不同取值情况下的条件分布是许多研究的主要目的。

2.1.2　灵敏度和准确度

表 2.2 的结果取自最近的一篇关于乳腺癌的不同诊断方式的文章。该文章回顾了有关文献,给出了使用乳房 X 光片与临床检查相结合的技术对诊断结果的影响。令 X = 真实的患病状况(即,妇女是否确实患有乳腺癌),Y = 诊断结果(阳性、阴性),其中阳性结果表示该妇女被诊断患有乳腺癌。表 2.2 中所估计的概率为在给定 X 时 Y 的条件概率。

表 2.2　估计的乳腺癌诊断结果的条件分布

乳腺癌	诊断结果		小　计
	阳性	阴性	
是	0.82	0.18	1.0
否	0.01	0.99	1.0

来源:数据取自 W. Lawrence et al., *J. Natl. Cancer Inst.* 90: 1792-1800 (1998)。

在诊断疾病时,两种正确的诊断结果分别为对患有疾病的研究对象诊断结果为阳性以及对没有患病的研究对象诊断结果为阴性。给定研究对象确实患有疾病,诊断结果为阳性的条件概率被称为*灵敏度*(*sensitivity*);给定研究对象没有患病,诊断结果为阴性的条件概率被称为*准确度*(*specificity*)(Yerushalmy, 1947)。理想情况下,两者的值都应该很高。

对于具有表 2.2 形式的 2×2 表格,灵敏度等于 $\pi_{1|1}$,准确度为 $\pi_{2|2}$。在表 2.2 中,结合使用乳房 X 光片与临床检查所估计的灵敏度为 0.82,即在患有乳腺癌的妇女中,82% 的患者可以得到正确的诊断。所估计的该方法的准确度为 0.99,即在未患有乳腺癌的妇女中,99% 可以得到正确诊断。

2.1.3 分类变量的独立性

当两个变量都是结果变量时,它们的关联可以通过二者的联合分布、给定 X 时 Y 的条件分布或者给定 Y 时 X 的条件分布来描述。给定 X 时 Y 的条件分布与联合分布的关系可表示为:

$$\text{对于所有的 } i \text{ 和 } j, \quad \pi_{j|i} = \pi_{ij}/\pi_{i+}。$$

如果两个分类结果变量的联合分布概率都等于相应的边际分布概率的乘积,那么我们就称这两个变量*相互独立*(*independent*),

$$\text{对于 } i = 1,\cdots,I \text{ 且 } j = 1,\cdots,J, \qquad \pi_{ij} = \pi_{i+}\pi_{+j}。\tag{2.1}$$

当 X 和 Y 相互独立时,

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\pi_{+j})/\pi_{i+} = \pi_{+j}, \qquad \text{其中 } i = 1,\cdots,I。$$

Y 的每一个条件分布与它的边际分布相同。因此,当$\{\pi_{j|1} = \cdots = \pi_{j|I}$,其中 $j = 1,\cdots,J\}$时,两个变量相互独立,也即,在每一行中结果变量落在任何列的概率都是相同的。以 Y 为结果变量,X 为解释变量,比公式 2.1 更自然的关于相互独立的定义为:相互独立是指条件分布的*同质性*(*homogeneity*)。

表 2.3 给出了 2×2 表格中关于联合分布,条件分布,以及边际分布的表达符号。描述样本分布的表达符号与此相似,只不过用 p 或 $\hat{\pi}$ 取代 π。例如,$\{p_{ij}\}$代表样本的联合分布。$\{n_{ij}\}$表示表中单元格的频数,并且 $n = \sum_i \sum_j n_{ij}$是总的样本规模。因而,

$$p_{ij} = n_{ij}/n。$$

在第 i 行的研究对象中落在第 j 列的样本比例等于

$$p_{j|i} = p_{ij}/p_{i+} = n_{ij}/n_{i+},$$

其中 $n_{i+} = np_{i+} = \sum_j n_{ij}$。

表 2.3 联合分布、条件分布和边际分布概率的表达符号

行	列		小 计
	1	2	
1	$\pi_{11}(\pi_{1|1})$	$\pi_{12}(\pi_{2|1})$	$\pi_{1+}(1.0)$
2	$\pi_{21}(\pi_{1|2})$	$\pi_{22}(\pi_{2|2})$	$\pi_{2+}(1.0)$
小计	π_{+1}	π_{+2}	1.0

2.1.4 泊松抽样、二项抽样和多项抽样

在第 1.2 节介绍的概率分布可以扩展到列联表中的单元格计数。例如,泊松抽样模型将单元格计数$\{Y_{ij}\}$视为参数为$\{\mu_{ij}\}$的独立泊松随机变量。这时,可能结果$\{n_{ij}\}$的联合概率密度函数等于 IJ 个单元格的泊松分布概率 $P(Y_{ij} = n_{ij})$的乘积,也即

$$\prod_i \prod_j \exp(-\mu_{ij})\mu_{ij}^{n_{ij}}/n_{ij}!。$$

当总的样本规模 n 是固定的但行和列的总计可变时,可以考虑*多项抽样*(*multinomial sampling*)模型。这时,IJ 个单元格代表可能的不同结果。单元格计数的概率密度函数具有多项分布形式

$$\left[n!/(n_{11}!\cdots n_{IJ}!)\right]\prod_i \prod_j \pi_{ij}^{n_{ij}}。$$

通常来说,对结果变量 Y 的观测是根据解释变量 X 的不同取值而分别进行的。这种

情况下,一般假定行总计给定,为了简便起见,我们使用符号 $n_i = n_{i+}$ 来表示。假定在 X 取 i 值时,关于 Y 的 n_i 个观测值是相互独立的,每个观测值的概率分布为 $\{\pi_{1|i},\cdots,\pi_{J|i}\}$,那么满足 $\sum_j n_{ij} = n_i$ 的计数 $\{n_{ij}, j=1,\cdots,J\}$ 具有多项分布形式

$$\frac{n_i!}{\prod_j n_{ij}!}\prod_j \pi_{j|i}^{n_{ij}}。\tag{2.2}$$

当 X 取不同值,对应的样本相互独立时,整个数据的联合概率函数等于多项分布函数(式 2.2)在不同 X 取值下的乘积。这种抽样设计称为*独立多项抽样*(*independent multinomial sampling*)或者*乘积多项抽样*(*product multinomial sampling*)。

独立多项抽样也可以在以下情况下发生:假定 $\{n_{ij}\}$ 来自于均值为 $\{\mu_{ij}\}$ 的独立泊松抽样或概率为 $\{\pi_{ij}=\mu_{ij}/n\}$ 的 IJ 个单元格的多项抽样。当 X 是解释变量时,即使按照抽样设计中其总数 $\{n_i = \sum_j n_{ij}\}$ 不固定,按照其总数为给定的情况进行统计推断也是合理的。在 $\{n_i\}$ 给定的条件下,单元格计数 $\{n_{ij}, j=1,\cdots,J\}$ 服从概率为 $\{\pi_{j|i}=\mu_{ij}/\mu_{i+}, j=1,\cdots,J\}$ 的多项分布(式 2.2),并且不同行中的单元格计数是相互独立的。在这种情况下,我们在分析数据中将行总计视作固定的,认为数据由分别的独立抽样形成。

有时行和列的边际是自然给定的。适合这种情况的抽样分布是*超几何分布*(*hypergeometric*)。这种情况较为少见,我们将在第 3.5.1 节讨论。

2.1.5 例子:安全带的使用

马萨诸塞州(Massachusetts)高速公路局的研究人员计划考察在马萨诸塞收费公路(turnpike)上发生的车祸中司机是否使用安全带(是、否)与事故结果(死亡、非死亡)之间的关系。他们准备将结果表示为表 2.4 的格式,计划把下一年在收费公路上发生的事故按照这些变量加以划分汇总。这样,总的样本规模是一个随机变量。他们可以将安全带使用情况与事故结果的四种组合的观测值计数视为相互独立的泊松随机变量,具有未知的均值 $\{\mu_{11},\mu_{12},\mu_{21},\mu_{22}\}$。

表 2.4 是否使用安全带与车祸结果

是否使用安全带	事故结果	
	死 亡	未死亡
是		
否		

另一种情况,假设研究者随机抽取了 200 份去年在收费公路上发生的车祸的警方记录,并按照是否使用安全带和事故结果进行了划分。在此研究中,总的样本规模 n 是给定的。那么,他们可以将四个单元格的计数作为一个 $n=200$ 次试验的多项分布随机变量,具有未知的联合分布 $\{\pi_{11},\pi_{12},\pi_{21},\pi_{22}\}$。

最后,假设出现死亡的警方记录与其他记录是分别归档的。研究者或许因而考虑随机选取 100 份出现死亡的事故记录,再随机选取 100 份未出现死亡的事故记录。这种方法固定了表 2.4 中的列总计为 100。这时,他们可以将表 2.4 中的每列当作一个独立的二项分布样本。还有一种情况,即传统的实验设计:选取 200 个研究对象,从中随机分配 100 个系安全带;接着强制 200 个研究对象全都经历一次交通事故。这时,结果中的每一行是独立的二项分布样本,行总计均为固定的 100(显然,在其他实验科学中常用的传统设计无法直接应用于对人类的研究。这一点在医学研究中尤为突出)。

2.1.6 研究的类型

表2.5所示为 Richard Doll 和 A. Bradford Hill 最早做的关于肺癌与吸烟之间关系的研究。在英国伦敦的20所医院里,对前一年因肺癌而住院的病人询问了他们的吸烟行为。研究者调查了在同一医院里与这709名肺癌住院患者具有相同性别、年龄(5岁组)的709名非癌症病人。表2.5中第一列为709名肺癌患者作为个案(*cases*),第二列为709名非癌症患者作为*控制案例*(*controls*)。在该研究中,吸烟者是指每天至少吸一支烟并持续了至少一年的人。

表2.5 关于肺癌与吸烟的交叉列联表

吸烟者	肺 癌	
	个 案	控制案例
是	688	650
否	21	59
小计	709	709

来源: Based on data reported in Table IV, R. Doll and A. B. Hill, *British Med. J.*, Sept. 30, 1950, pp. 739-748.

一般来说,是否得了肺癌应该是结果变量,而吸烟行为是解释变量。然而,在这项研究中,按照抽样设计,肺癌的边际分布是给定的,而所度量的结果是研究对象是否曾经吸过烟。该类研究,通过使用*回顾性*(*retrospective*)设计"考察过去",称为*个案-控制研究*(*case-control study*)。这样的研究在与健康有关的领域内应用非常广泛。通常,与上述研究相同,两个样本通过匹配来产生。也有时候,个案样本和控制样本是相互独立的而不是匹配的。例如,早期另一项对肺癌和吸烟的个案-控制研究,通过向1950或1951年死于某种癌症的医师的地址寄信来抽取调查对象,并将所得到的观测值按照癌症类别以及研究对象的吸烟行为进行了交叉划分(参见,如 Cornfield 1956)。

事实上我们想比较吸烟者和非吸烟者中患有肺癌的比例。这些比例是在给定吸烟行为下发生肺癌的条件分布。与之相反,个案-控制研究给出的是相反方向的比例,即在给定肺癌状况的情况下吸烟行为的条件分布。在表2.5患有肺癌的人中,曾经是吸烟者的比例为688/709 = 0.970,而在控制样本中该比例为650/709 = 0.917。

当我们知道总体中患有肺癌的比例时,就可以通过贝叶斯定理(Bayes' theorem)去计算我们感兴趣的问题对应的样本条件分布(习题2.21)。否则,通过一个回顾性的样本我们无法估计在每一类的吸烟行为中发生肺癌的比例。对于表2.5,我们无法知道总体中肺癌的患病率,调查对象的发病率可能比一般人群高得多。

相反,设想一项研究从青少年群体中选取样本,然后过60年后再观察吸烟者与非吸烟者中罹患肺癌的比率。这样的抽样设计是*前瞻性的*(*prospective*)。一般前瞻性研究可以分为两种。*临床试验*(*clinical trials*)是将研究对象随机分配为吸烟者和非吸烟者两组。在*队列研究*(*cohort studies*)中,研究对象自主决定是否吸烟,然后观察其未来肺癌发生的情况。还有另外一种方法,就是*横截面设计*(*cross-sectional design*),选取调查对象并将其同时按照两个变量进行划分。

前瞻性研究(*Prospective studies*)通常在给定 X 的各类别总计 $\{n_i = \sum_j n_{ij}\}$ 的条件下,将每行中的 J 个计数视为一个关于 Y 的独立多项分布样本。*回顾性研究*(*Retrospective studies*)通常将 Y 的总计 $\{n_{+j}\}$ 当作给定的,并将每列中的 I 个计数作为一个关于 X 的独立多项分布样本。在*横截面研究*(*cross-sectional studies*)中,总的样本规模是给定的,但是

行总计或列总计是可变的，这时 IJ 个单元格计数是一个多项分布样本。

个案-控制研究、队列研究，以及横截面研究统称为*观察研究*（*observational studies*）。它们仅仅观察谁选择哪一组以及谁发生了研究所关注的结果。相反，临床试验是一种实验（*experimental*）研究，研究者可以决定哪个研究对象得到哪种方式的治疗。这样的研究可以通过随机分配样本确保各组中与结果变量相关的其他变量的分布大致相当。相对而言，观察研究更为普遍，但它也更容易受到各种偏差的影响。

2.2 两个比例的比较

许多研究试图在不同组间比较一个二分结果变量，这时，Y 仅包括两个类别，比如一种医疗方式的结果（成功，失败）。在只存在两组的情况下，可以通过 2×2 的列联表来表示研究结果。其中，列联表的行表示组别，列则表示结果变量 Y 的类别。我们在这一节介绍进行组间比较的参数。

2.2.1 比例之差

对于第 i 行中的研究对象，$\pi_{1|i}$是结果落在第 1 类（“成功”）的概率。由于仅存在两种可能的结果，$\pi_{2|i}=1-\pi_{1|i}$。为了简便起见，我们使用符号 π_i 代表 $\pi_{1|i}$。成功所占的*比例之差*（*difference of proportions*），即 $\pi_1-\pi_2$，是对两行之间结果的一个基本比较。用失败所占的比例进行比较是等价的，因为

$$(1-\pi_1)-(1-\pi_2)=\pi_2-\pi_1。$$

比例之差的取值范围在 -1.0 到 $+1.0$ 之间。如果各行具有完全相同的条件分布，则比例之差等于零。当 $\pi_1-\pi_2=0$ 时，结果变量 Y 在统计上独立于行变量。

如果两个变量都是结果变量，可以讨论在任一方向上的条件分布。这时，也可以对两列进行比较，如各列中第 1 行所占的比例之差。在两列之间进行比较的结果往往并不等于两行之间的比例之差 $\pi_1-\pi_2$。

2.2.2 相对风险

对于同样大小的 $\pi_1-\pi_2$ 的值，当两个 π_i 都接近于 0 或 1 时比在 π_i 取其他值时更加重要。例如，比较两种不同治疗方式对研究对象的死亡比例的影响，0.010 与 0.001 之间的差或许比 0.410 与 0.401 之间的差更值得关注，尽管它们都等于 0.009。在这种情况下，比例之比提供了重要信息。

相对风险（*relative risk*）被定义为

$$\pi_1/\pi_2。\tag{2.3}$$

它可以取任意非负的实数值。相对风险为 1.0 对应着相互独立的情况。就上文所给的比例而言，相对风险分别等于 $0.010/0.001=10.0$ 和 $0.410/0.401=1.02$。若按照结果变量的第二个类别进行行间的比较，会得出不同的相对风险，即 $(1-\pi_1)/(1-\pi_2)$。

2.2.3 发生比之比

在事件成功的概率为 π 时，*发生比*（*odds*）被定义为

$$\Omega=\pi/(1-\pi)。$$

发生比的取值不能为负，当成功的结果比失败更可能发生时，$\Omega>1.0$。例如，当 $\pi=0.75$

时，$\Omega = 0.75/0.25 = 3.0$；成功发生的可能性是失败发生的可能性的三倍，每观察到一次失败结果，我们期望观察到大约三次成功。当 $\Omega = \frac{1}{3}$ 时，失败发生的可能性是成功发生的可能性的三倍。相应地，

$$\pi = \Omega/(\Omega + 1)。$$

例如，当 $\Omega = \frac{1}{3}$ 时，$\pi = 0.25$。

在 2×2 表格中，在第 i 行的成功相对于失败的发生比是 $\Omega_i = \pi_i/(1 - \pi_i)$。两行的发生比 Ω_1 与 Ω_2 之间的比

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} \tag{2.4}$$

被称为*发生比之比*（*odds ratio*）。

对于单元格概率为 $\{\pi_{ij}\}$ 的联合分布，第 i 行的发生比可以等价地定义为 $\Omega_i = \pi_{i1}/\pi_{i2}$，$i = 1,2$。那么发生比之比等于

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}。 \tag{2.5}$$

θ 也被称为*交叉乘积比*（*cross-product ratio*），因为它等于两个相反对角线上的概率的乘积 $\pi_{11}\pi_{22}$ 和 $\pi_{12}\pi_{21}$ 之比（Yule, 1900, 1912）。

2.2.4 发生比之比的特性

发生比之比可以取任意非负的值。如果 $\Omega_1 = \Omega_2$，从而（当所有的单元格概率为正时）$\theta = 1$，那么 X 和 Y 相互独立。当 $1 < \theta < \infty$ 时，第 1 行的研究对象比第 2 行的研究对象结果为成功的可能性更大，也即 $\pi_1 > \pi_2$。例如，当 $\theta = 4$ 时，第 1 行结果为成功的发生比是第 2 行的四倍。这并不意味着*概率* $\pi_1 = 4\pi_2$，它对应的是*相对风险*（*relative risk*）等于 4.0时的情况。当 $0 < \theta < 1$ 时，$\pi_1 < \pi_2$。如果某个单元格的概率为零，那么 θ 等于 0 或 ∞。

θ 的值在任一方向上与 1.0 相差越大表示关联的强度也越大。在 1.0 的两侧，当一个 θ 值是另一个的倒数时，这两个值代表相同的关联。例如，当 $\theta = 0.25$ 时，第 1 行成功的发生比是第 2 行的 0.25 倍，或者等价地说，第 2 行成功的发生比是第 1 行的 $1/0.25 = 4.0$ 倍。当行的顺序或列的顺序互换后，新的 θ 值等于初始值的倒数。

在统计推断中，将看到使用 $\log\theta$ 更加方便。当两个变量相互独立时，$\log\theta = 0$。对数发生比之比以此为中点呈对称分布——改变行或列的位置会导致它的符号发生变化。两个符号相反、绝对值相同的 $\log\theta$，如 $\log 4 = 1.39$ 和 $\log 0.25 = -1.39$，代表着同等强度的关联。

当表格的方位发生改变时，即原来的行变成列而原来的列变成行时，发生比之比的值保持不变。这一结果从公式 2.5 的对称性很容易得出。在使用 θ 时，没有必要区分哪个变量是结果变量。实际上，尽管公式 2.4 通过有关 $\pi_i = P(Y = 1 \mid X = i)$ 的发生比来定义 θ，使用相反的条件概率进行定义，结果完全相同。当每个方向上的联合分布、条件分布都存在时，那么

$$\begin{aligned} \theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} &= \frac{P(Y = 1 \mid X = 1)/P(Y = 2 \mid X = 1)}{P(Y = 1 \mid X = 2)/P(Y = 2 \mid X = 2)} \\ &= \frac{P(X = 1 \mid Y = 1)/P(X = 2 \mid Y = 1)}{P(X = 1 \mid Y = 2)/P(X = 2 \mid Y = 2)}。 \end{aligned} \tag{2.6}$$

事实上,无论是前瞻性、回顾性,还是横截面抽样设计,发生比之比都同样适用。无论在哪种情况下,样本的发生比之比估计的都是同一个参数。

对于单元格计数$\{n_{ij}\}$而言,样本的发生比之比为

$$\hat{\theta} = n_{11}n_{22}/n_{12}n_{21}。$$

当任一行或列中的两个单元格计数都乘以一个非零的常数时,它的值保持不变。这一特性的含义是,即使一个变量的不同类别的样本是按照很大或很小的不等比例抽样获得的,样本发生比之比所估计的仍然是同一个参数(θ)。举例来说,一项有关注射疫苗与感染某种流感的关系的回顾性研究中,在以下两种随机样本的情况下,样本发生比之比估计的是同一个值:(1)100 个得过流感的与 100 个未得过流感的人,(2)40 个得过流感的与 160 个未得过流感的人。对于同一行内的每个计数乘上一个常数来说,样本的比例之差和相对风险(式 2.3)也不会发生变化,但是当同一列内的每个计数乘上一个常数或行和列交换位置时,它们就会发生变化。

2.2.5 例子:再论阿司匹林与心脏病的关系

我们通过表 2.1 中关于服用阿司匹林与发生心脏病的数据来展示这三种度量关联的指标。该表中区分了致命性和非致命性心脏病,在此我们将这两种结果加以合并。在 11 034 位服用安慰剂的医师中,189 位患有心脏病,比例为 189/11 034 = 0.017 1。在 11 037位服用阿司匹林的医师中,104 位出现了心脏病,比例为 0.009 4。样本的比例之差为 0.017 1 − 0.009 4 = 0.007 7。样本的相对风险为 0.017 1/0.009 4 = 1.82。服用安慰剂的医师得心脏病的比例是服用阿司匹林的 1.82 倍。样本的发生比之比为(189 × 10 933)/(10 845 × 104) = 1.83。服用安慰剂的医师患心脏病的发生比是服用阿司匹林的相应发生比的 1.83 倍。

2.2.6 个案-控制研究与发生比之比

在回顾性抽样设计中,如个案-控制研究,我们可以估计如$P(X=i|Y=j)$的条件概率。这种情况下,通常我们无法估计针对所关注的结果的发生概率$P(Y=j|X=i)$或者相应的比例之差以及相对风险。但是,我们仍然可以估计发生比之比,因为通过公式 2.6 可知,发生比之比可由任一方向上的条件概率决定。

举例来说,我们回到表 2.5,其中X= 吸烟行为,Y= 肺癌。该数据是在给定Y的取值上的关于X的两个二项分布样本。我们可以估计在给定研究对象是否患有肺癌时他曾经抽烟的概率;在个案(cases)一组中该概率为 688/709,在控制(controls)一组中它等于 650/709。但这时我们无法估计给定一个人是否曾经吸烟的情况下他患肺癌的概率,而这才是我们所关注的重点。我们也无法估计患肺癌的概率的差或比。比例之差和相对风险仅仅局限于比较是否曾经吸烟的概率。然而,根据公式 2.6,我们可以计算这个样本的发生比之比,

$$\frac{(688/709)(21/709)}{(650/709)(59/709)} = \frac{688 \times 59}{650 \times 21} = 3.0。$$

由公式 2.6 可知,即便在回顾性研究中,仍然可以按照所关注的方向来解释结果:所估计的吸烟者患肺癌的发生比是非吸烟者的相应发生比的 3.0 倍。

2.2.7 发生比之比与相对风险的关系

由定义公式 2.3 和 2.4 可得,

$$发生比之比 = 相对风险 \times \frac{1-\pi_2}{1-\pi_1}。$$

当所关注结果的概率 π_i 在比较的两组中都接近于零时，它们的大小相当。我们在第2.2.5节关于阿司匹林的研究中已经看到了这种相似性，其中每一组患心脏病的比例都小于0.02。它的相对风险为1.82，发生比之比为1.83。

由于这种相似性，当每个 π_i 都很小时，在无法直接估计相对风险情况下，如在个案-控制研究中，发生比之比提供了一个对相对风险的粗略估测（Cornfield, 1951）。例如，在表2.5中，如果无论吸烟与否，罹患肺癌的概率都很低，3.0就可以作为对相对风险的一个粗略估计；也即，吸烟者患肺癌的相对频率大约是非吸烟者的3.0倍。

2.3 分层 2×2 表格中的偏关联

在许多研究中，尤其是观察研究中，一个非常重要的步骤是对控制变量的选取。在研究 X 对 Y 的影响时，需要控制能够影响两者之间关系的协变量（covariate）。这就要求运用某种方法保证协变量的值恒定。否则，所观察到的 X 对 Y 的效应可能实际上反映的是协变量同时对 X 和 Y 的影响。这时 X 和 Y 的相关关系出现了*混淆*（*confounding*）问题。实验研究可以通过将研究对象按照 X 的不同取值随机分配而消除混淆协变量的影响，但是在观察研究中我们无法做到这一点。

假设一项研究考察被动吸烟的影响，即与一名吸烟者同住对非吸烟者的影响。分析被动吸烟是否与患肺癌相关，横截面研究可能会比较配偶吸烟的非吸烟者与配偶不吸烟的非吸烟者的肺癌发病率。这样的研究应当考虑控制年龄、社会经济地位或其他可能同时影响配偶吸烟和患肺癌风险的因素。否则，研究结果的价值有限：配偶不吸烟的人或许比配偶吸烟的人更年轻，而年轻人患肺癌的可能性较低。这时，配偶不吸烟的人肺癌发生的比例比配偶吸烟的人低可能仅仅反映了前者平均年龄较小的影响。

在本节，我们讨论当控制了一个可能的混淆变量 Z 后，对分类变量 X 和 Y 的关联的分析。为了简便起见，我们先考虑只存在一个控制变量的情况。在后面的章节中，我们将介绍更一般的情况并讨论通过模型来实现统计控制。

2.3.1 分表

在研究 X 和 Y 的关联时，我们通过固定 Z 的取值来对其进行控制。三维列联表在 Z 的每一个取值下对应一组关于 X 和 Y 的二维交叉列联表。这些二维表被称为*分表*（*partial tables*）。它们展示的是通过限定 Z 的值不变而消除其影响后的 XY 关联。

对分表进行合并得到的二维列联表被称为 *XY 边际表*（*marginal table*）。边际表中的每个单元格计数等于对所有分表中相同位置的单元格计数的加总。边际表没有控制 Z，而是忽略了它。边际表本身不包含关于 Z 的任何信息，它仅仅是一个关于 X 和 Y 关系的二维表，其显示的 XY 关系有可能是 Z 对 X 和 Y 的影响。

在分表中存在的关联被称为*条件关联*（*conditional associations*），即它们指的是在给定 Z 取某一值的条件下 X 对 Y 的影响。分表中的条件关联可能会与边际表中的关联存在很大差别。事实上，仅仅分析一个多维列联表中的边际表往往会得出误导性的结论。下文提供了一个很好的例子。

2.3.2 例子:死刑判决

表2.6所示为种族特征对谋杀犯是否会被判死刑的影响的2×2×2列联表——包括两行、两列和两层。该表中的674名研究对象为佛罗里达州1976—1987年间因卷入多起谋杀案件而被控的被告。所包括的变量为$Y=$判处死刑,分为(是、否)两类,$X=$被告的种族,以及$Z=$受害人的种族,类别为(白人、黑人)两种。我们分析被告人种族对死刑判决结果的影响,将受害人的种族视为一个控制变量。对于受害人种族的每一取值,表2.6包含一个关于被告人种族与死刑判决结果的2×2分表。

表2.6 按照被告人和受害人种族划分的死刑判决结果

受害人种族	被告人种族	死刑		获判死刑的百分比
		是	否	
白人	白人	53	414	11.3
	黑人	11	37	22.9
黑人	白人	0	16	0.0
	黑人	4	139	2.8
小计	白人	53	430	11.0
	黑人	15	176	7.9

来源:M. L. Radelet and G. L. Pierce, *Florida Law Rev.* 43: 1-34 (1991). Reprinted with permission from the *Florida Law Review*.

对于被告人和受害人种族的每一组合,表2.6和图2.1给出了被告人获判死刑的百分比,这些数字是对条件关联的描述。当被害人为白人时,黑人被告获判死刑的百分比比白人被告高22.9% -11.3% =11.6%。当被害人是黑人时,黑人被告获判死刑的百分比比白人被告高2.8%。给定被害人种族不变来对其进行*控制*,与白人被告相比,黑人被告获判死刑的情况更普遍。

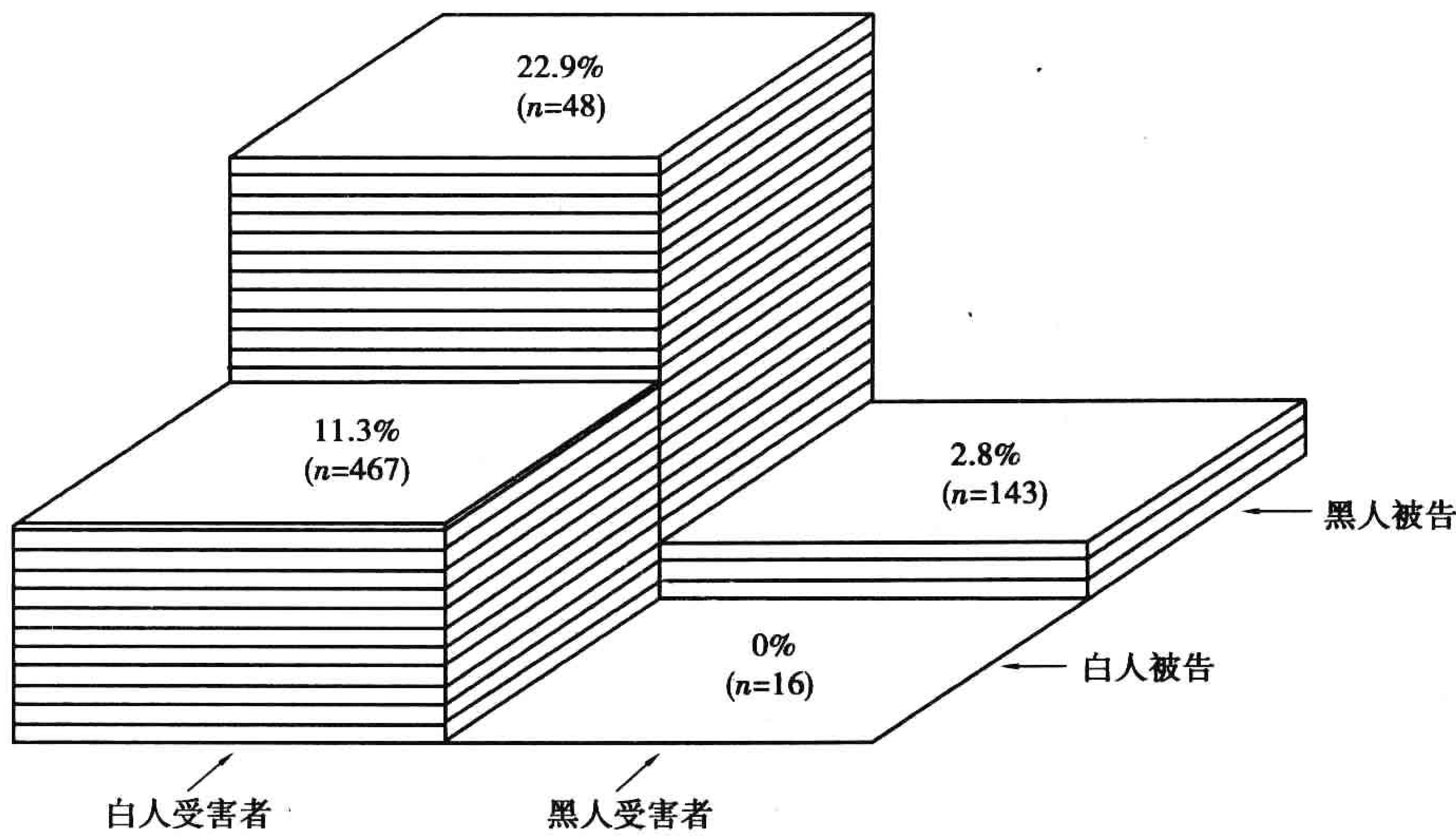


图2.1 获判死刑的百分比

表2.6的底部给出了边际表。它相当于对表2.6中不同种族被害人的两组单元格计数加总,也即合并了两个分表(如11+4=15)。总体而言,11.0%的白人被告和7.9%的黑人被告被判决了死刑。*忽略掉*受害人的种族,黑人被告被判死刑的比例比白人被告还小。这个结果与分表的结果正好相反。

为什么在我们忽略被害人的种族与对其进行统计控制时相比,被告人种族与获判死刑的关联会出现如此大的变化?这取决于被害人种族与其他两个变量的关联。首先,被害人种族与被告人种族之间的关联非常强。由这两个变量所形成的边际表的发生比之比等于$(467\times143)/(48\times16)=87.0$。其次,表2.6表明,不论被告人的种族,当被害人为白人而不是黑人时,被告更有可能被判决死刑。因此,白人倾向于杀害白人,而杀害白人更容易招致死刑。这就导致与条件关联相比,边际关联显示白人被告更容易被判死刑的假象。事实上,表2.6印证了这一点。

图2.2展示了为什么边际关联会不同于条件关联。对于被告人的不同种族,该图画出了按照受害人种族划分的判决死刑的比例。图中用字母注明了受害人的种族,并用大小不等的圈来表示按照被告人和受害人种族所划分的观测值占总样本的比例。例如,最大的圈里的 *W* 表示当被告人和受害人都是白人时被判死刑的比例为0.113。由于落在这一组合的观测值最多$(53+414=467)$,所以这个圈最大。类似地,第二大的圈表示被告人和受害人均为黑人的情形。

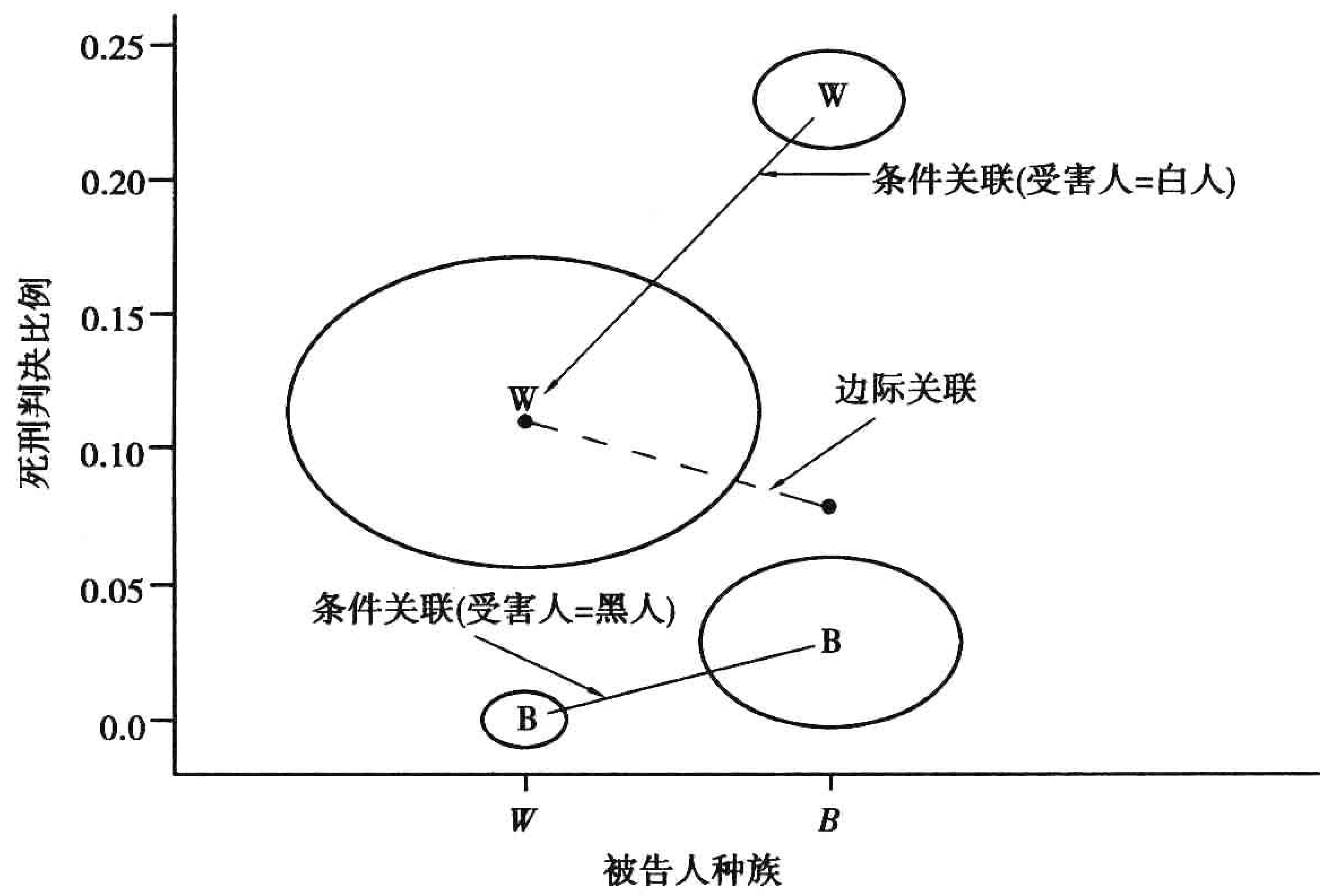


图2.2　在控制或忽略被害人种族情况下按照被告人种族划分的死刑判决比例

通过比较受害人种族相同的圈(中心的字母一样),我们对其进行了控制。连接两个 *W* 的线斜率为正,连接两个 *B* 的线也是如此。这表明在控制了受害人种族后,黑人被告与白人被告相比更可能被判死刑。当我们将不同受害人种族的结果合并时,得到了被告人种族对死刑判决的边际影响。这时,较大的圈由于具有较多的样本量而具有更大的影响力。因而,不同种族被告人的加总比例,在图中由圆点来表示,落在离大圈的中心较近的地方。连接边际比例的线斜率为负,表明总体上白人被告比黑人被告更可能被判死刑。

这种边际关联的结果与每个条件关联的结果方向相反的情况被称为*辛普森悖论*(*Simpson's paradox*)(Simpson, 1951; Yule, 1903)。它既适用于定量变量也适用于分类变量。统计学家经常用它来警告从 *X* 和 *Y* 的关联来推论因果关系的危险性。例如,当医生开始观察到吸烟与肺癌之间存在很强的发生比之比时,诸如 R. A. Fisher 等统计学家们则强调,可能存在其他变量(如基因因素)会使在进行相应控制的情况下吸烟与肺癌之间的关联消失。然而,其他统计学家(如 J. Cornfield)指出,在 *XY* 存在强关联的情况下,混淆变量 *Z* 必须与 *X* 和 *Y* 都存在非常强的关联,才有可能使得在控制 *Z* 的条件下 *X* 对 *Y* 的影响消失或发生变化(Breslow and Day, 1980, Sec. 3.4)。

2.3.3 条件发生比之比与边际发生比之比

发生比之比既可以用来描述边际关联也可以用来描述条件关联。我们讨论一个 $2\times2\times K$ 表格的情况，这里 K 表示控制变量 Z 的可选类别数。令 $\{\mu_{ijk}\}$ 表示在某一抽样模型下——如二项分布、多项分布或泊松分布抽样——的单元格期望频数。

给定 Z 的取值为类别 k，则发生比之比为

$$\theta_{XY(k)}=\frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} \tag{2.7}$$

它描述在分表 k 中的 XY 条件关联的强度。这 K 个分表的发生比之比被称为*条件发生比之比*(*conditional odds ratios*)。条件发生比之比可能与边际发生比之比存在很大的差异。XY 的边际表具有期望频数 $\{\mu_{ij+}=\sum_k\mu_{ijk}\}$ 。因此，XY 的边际发生比之比等于

$$\theta_{XY}=\frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}。$$

计算 $\theta_{XY(k)}$ 和 θ_{XY} 的样本统计量的公式与上式相同，只是用单元格计数取代了期望频数。我们用表 2.6 中被告人种族与死刑判决的关联来对此加以说明。在第一个分表中，受害人的种族为白人，有

$$\hat{\theta}_{XY(1)}=\frac{53\times37}{414\times11}=0.43。$$

白人被告获判死刑的样本发生比是黑人被告的样本发生比的 43%。在第二个分表中，受害人为黑人时估计的发生比之比为 $\hat{\theta}_{XY(2)}=(0\times139)/(16\times4)=0.0$，因为当受害人为黑人时，从来没有白人被告被判过死刑。

将受害人的种族进行合并，根据表 2.6 中的 2×2 边际表来估计边际发生比之比，有 $(53\times176)/(430\times15)=1.45$。白人被告获判死刑的样本发生比比黑人被告高 45%。然而，给定受害人的种族，白人被告的发生比之比却较小。这种在控制了受害人种族后出现的关联方向改变的现象就是辛普森悖论。

2.3.4 边际独立与条件独立

更一般地，X 可能包括 I 个类别，Y 可能包括 J 个类别。一个 $I\times J\times K$ 表格描述在控制了 Z 的条件下 X 和 Y 的关系。如果 X 和 Y 在分表 k 中相互独立，那么 X 和 Y 被称为在 Z 的取值为 k 时*条件独立*(*conditional independence*)。当 Y 是结果变量时，这意味着

$$\text{对所有 } i,j,\quad P(Y=j\mid X=i,Z=k)=P(Y=j\mid Z=k)。 \tag{2.8}$$

进一步地，如果 X 和 Y 在 Z 的任何一个取值上都条件独立，那么可以说 X 和 Y 在给定 Z 时*条件独立*(*conditionally independent given Z*)。也即，当式 2.8 对于所有 k 都成立时，那么，给定 Z，Y 不依赖于 X。

假定整个三维表格满足一个多项分布，其联合概率为 $\{\pi_{ijk}=P(X=i,Y=j,Z=k)\}$。这时

$$\pi_{ijk}=P(X=i,Z=k)P(Y=j\mid X=i,Z=k),$$

给定 Z，X 和 Y 条件独立时，上式等于

$$\pi_{i+k}P(Y=j\mid Z=k)=\pi_{i+k}P(Y=j,Z=k)/P(Z=k)。$$

因此，条件独立等价于

$$\text{对所有 } i,j,k,\pi_{ijk}=\pi_{i+k}\pi_{+jk}/\pi_{++k}。 \tag{2.9}$$

条件独立并不意味着边际独立（Yule, 1903）。例如，将式 2.9 的两边都对 k 加总得到

$$\pi_{ij+} = \sum_k (\pi_{i+k}\pi_{+jk}/\pi_{++k})。$$

所有需要加总的三项都包含 k，它无法简化为 $\pi_{ij+} = \pi_{i++}\pi_{+j+}$，即边际独立。

对于 $2 \times 2 \times K$ 表格，当 X 和 Y 的发生比之比在 Z 的每个取值上都等于 1 时，X 和 Y 条件独立。表 2.7 中的期望频数 $\{\mu_{ijk}\}$ 展示了这种关系，其中 $Y=$ 治疗结果（成功，失败），$X=$ 治疗方式（A，B），以及 $Z=$ 诊所（1，2）。由式 2.7 可知，XY 的条件发生比之比为

$$\theta_{XY(1)} = \frac{18 \times 8}{12 \times 12} = 1.0, \quad \theta_{XY(2)} = \frac{2 \times 32}{8 \times 8} = 1.0。$$

给定每个诊所，治疗结果与治疗方式之间条件独立。对两个诊所的分表进行合并得到相应的边际表。边际发生比之比为 $\theta_{XY} = (20 \times 40)/(20 \times 20) = 2.0$，所以两个变量不满足边际独立。

表 2.7 条件独立不等于边际独立的期望频数

诊 所	治疗方式	结果	
		成 功	失 败
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
小计	A	20	20
	B	20	40

忽略掉诊所后，为什么治疗方式 A 结果为成功的发生比是治疗方式 B 的两倍？XZ 和 YZ 的条件发生比之比给出了原因。Z 与 X 或 Y 的发生比之比，在另一个变量取值给定的情况下，都等于 6.0。例如，在 Y 的取值为第一类的条件下，XZ 的发生比之比为 $(18 \times 8)/(12 \times 2) = 6.0$。在诊所 1 接受的治疗为方式 A 的条件发生比（给定结果）是诊所 2 的 6 倍，并且在诊所 1 治疗结果为成功的条件发生比（给定治疗方式）也是诊所 2 的 6 倍。诊所 1 更倾向于使用治疗方式 A，同时也出现了更多的成功结果。一种可能的情况是，如果诊所 1 的病人比诊所 2 的病人更年轻并且健康状况更好，或许不论采用哪种治疗方式，诊所 1 都会拥有较高的成功率。

仅仅分析边际表会导致错误的结论，认为治疗方式 A 的成功率更高。如上文所述，在某一特定诊所就医的对象有可能比总的样本更加同质，实际上在每个诊所治疗结果与治疗方式相互独立。

2.3.5 同质关联

在以下条件下，$2 \times 2 \times K$ 表格具有 XY 的*同质关联*（*homogeneousXY association*），

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}。$$

这时 X 对 Y 的影响在 Z 的每个取值上都是相同的。条件独立是当 $\theta_{XY(k)} = 1.0$ 时的一个特例。

在 XY 存在同质关联的情况下，其他的关联也满足同质性。例如，对于 Y 的每个取值，由 X 的两个类别与 Z 的任意两个类别所形成的条件发生比之比也完全相等。就发生

比之比来说，同质关联具有对称性。它适用于从第三个变量的不同类别来考察的任意两个变量。当同质关联存在时，可以说两个变量对第三个变量的影响*不存在交互效应*（*no interaction*）。

如果存在交互效应，任意两个变量的条件发生比之比会因第三个变量的不同取值而发生变化。如 $X=$ 吸烟（是，否）、$Y=$ 肺癌（是，否），以及 $Z=$ 年龄（<45，$45\sim65$，>65），假定 $\theta_{XY(1)}=1.2$，$\theta_{XY(2)}=3.9$，并且 $\theta_{XY(3)}=8.8$，那么对于年轻人而言吸烟对肺癌存在较小影响，但是随着年龄的增加这种影响明显加强。在这里，年龄被称为*效应修正因子*（*effect modifier*），吸烟的影响随着它的不同取值而发生变化。

就死刑判决的数据（表 2.6）而言，$\hat{\theta}_{XY(1)}=0.43$，$\hat{\theta}_{XY(2)}=0.0$，它们的值并不相近，但是由于存在计数为零的单元格（zero cell count），第二个估计并不稳定。将每个单元格计数都加上$\frac{1}{2}$后，$\hat{\theta}_{XY(2)}=0.94$。由于 $\hat{\theta}_{XY(2)}$ 不稳定，且存在抽样误差，这些分表的结果并不能说明在总体中一定不存在同质关联。在第 6.3 节我们将介绍如何分析样本数据是否满足同质关联或条件独立。

2.4 扩展到 $I\times J$ 表格

对于 2×2 表格，一个单一的数字（如发生比之比）便可以概括其中的关联。对于 $I\times J$表格，在不损失信息的情况下几乎不可能用一个数字来描述其中的关联。在这种情况下，往往需要通过一组发生比之比或其他综合指标来描述关联的某些特征。

2.4.1 $I\times J$ 表格的发生比之比

$I\times J$ 表格可以使用任意$\binom{I}{2}=I(I-1)/2$ 个两行和任意$\binom{J}{2}=J(J-1)/2$ 个两列的组合来计算发生比之比。以第 a,b 行与第 c,d 列为例，发生比之比$(\pi_{ac}\pi_{bd})/(\pi_{bc}\pi_{ad})$使用了矩形表格中的四个单元格，这样的发生比之比共有$\binom{I}{2}\binom{J}{2}$个。这些发生比之比中包含着很多冗余的信息。

考虑其中的一个子集，$(I-1)(J-1)$个*局部发生比之比*（*local odds ratios*）

$$\theta_{ij}=\frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}},\qquad i=1,\cdots,I-1,j=1,\cdots,J-1。\tag{2.10}$$

图 2.3 显示，局部发生比之比使用的是相邻行和相邻列的单元格。这$(I-1)(J-1)$个发生比之比决定了所有由任意两行和任意两列所对应的发生比之比。具体而言，在表 2.1 中，前两列的样本局部发生比之比为 2.08，第二和第三列的局部发生比之比为 1.74。无论在哪种情况下，都是在服用安慰剂的组较严重结果发生得更为普遍。这两个发生比之比的乘积等于 3.63，即是第一列和第三列的发生比之比。

像式 2.10 这样的非冗余发生比之比的最小子集并不唯一。另一组常用的子集为

$$\alpha_{ij}=\frac{\pi_{ij}\pi_{IJ}}{\pi_{Ij}\pi_{iJ}},\qquad i=1,\cdots,I-1,\ j=1,\cdots,J-1。\tag{2.11}$$

它使用了第 i 行第 j 列的单元格以及最后一行和最后一列的单元格所形成的矩形表。图 2.3 展示了这种情况。

给定边际分布$\{\pi_{i+}\}$和$\{\pi_{+j}\}$，当$\{\pi_{ij}>0\}$时，将概率转化为由式 2.10 或式 2.11 所

表示的一组发生比之比不会损失任何信息。单元格概率可以确定发生比之比,而在给定边际分布的情况下,发生比也可以确定单元格概率。从这个意义上,$(I-1)(J-1)$个参数能够描述$I \times J$表格中的任意关联。这时,独立性等价于所有$(I-1)(J-1)$个发生比之比等于1.0。

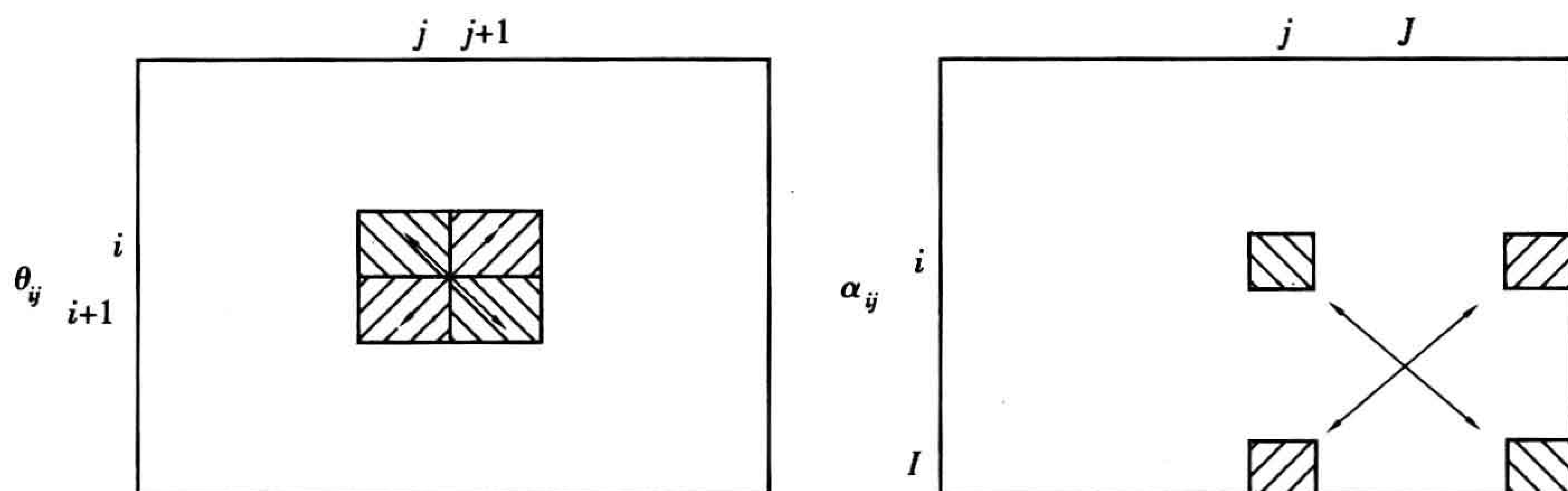


图2.3 $I \times J$表格的发生比之比

对于三维的$I \times J \times K$表格,分表中的一组发生比之比用来描述条件关联。XY同质关联意味着在给定Z的任一类别,由X的两个类别和Y的两个类别所形成的任何条件发生比之比都相等。

2.4.2 关联的总体度量指标

另一种描述关联的方法是使用一个单一的总体指标。我们先讨论定类变量的情形,再讨论定序变量。对于定类变量而言,最具有解释意义的指标与定距变量中的R^2结构相同。它和更为常见的组内相关系数(intraclass correlation coefficient)以及相关比(correlation ratio)(Kendall and Stuart, 1979)一样,描述从结果变量Y的边际分布到给定解释变量X后Y的条件分布的方差消减比例。

令$V(Y)$表示一个关于Y的边际分布$\{\pi_{+j}\}$的变异程度的度量指标,令$V(Y|i)$表示对在X取i值时Y的条件分布$\{\pi_{1|i}, \pi_{2|i}, \ldots, \pi_{J|i}\}$的变异程度的相应度量。变异消减比例的度量指标具有以下形式:

$$\frac{V(Y) - E[V(Y|X)]}{V(Y)}, \tag{2.12}$$

其中$E[V(Y|X)]$是在给定X的分布后Y的条件变异程度的期望值。X的边际分布为$\{\pi_{i+}\}$,则有$E[V(Y|X)] = \sum_i \pi_{i+} V(Y|i)$。

对于定类结果变量,Theil(1970)提出了一个度量变异程度的指标,即$V(Y) = \sum \pi_{+j} \cdot \log \pi_{+j}$,称为*熵*(*entropy*)。在列联表的情况下,熵的消减比例等于

$$U = -\frac{\sum_i \sum_j \pi_{ij} \log(\pi_{ij}/\pi_{i+}\pi_{+j})}{\sum_j \pi_{+j} \log \pi_{+j}}, \tag{2.13}$$

称为*不确定系数*(*uncertainty coefficient*)。当一个以上的$\pi_{+j} > 0$时,这个指标就可以确定。它的取值范围在0到1之间:$U=0$等价于X和Y相互独立;$U=1$等价于缺乏条件变异,即对于每一个i,都存在某个j使得$\pi_{j|i}=1$。

类似于式2.12描述$I \times J$表格中的关联的度量指标有很多(如习题2.38和2.39)。使用这些指标的一个难点在于,如何判断一个值到底多大才构成强关联。例如,当说熵的消减比例为30%时,它意味着什么?对于两个变量都是定序变量的情况,如下文所述,总体指标更易于解释,也更有价值。

2.4.3 定序趋势:相协对与相异对

表2.8中的变量为收入和工作满意度,数据来自一个关于美国黑人男性的全国性样本。两个变量的测度都是定序的,工作满意度的类别为非常不满意(VD)、有些不满意(LD)、基本满意(MS),以及非常满意(VS)。

当X和Y都是定序变量时,经常会存在一个单调的趋势关联。随着X取值的上升,对Y的回答也倾向于上升到较高的水平或者下降到较低的水平。例如,可能工作满意度会随着收入的增长而提高。可以用一个单一的参数来描述这种趋势,比如类似于相关系数的度量指标就可以描述这种单调关系。有些指标是基于对每一对研究对象是相协还是相异来进行划分的。如果两个对象中在X上排序较高者在Y上也排序较高,那么这一对被称为*相协*(*concordant*)。如果在X上排序较高的对象在Y上排序较低,那么这一对被称为*相异*(*discordant*)。当两个对象在X和/或Y上具有相同的取值时,那么这一对被称为*相平*(*tied*)。

表2.8 收入与工作满意度的交叉列联表

收入/美元	工作满意度			
	非常不满意	有点不满意	比较满意	非常满意
<15 000	1	3	10	6
15 000~25 000	2	3	10	7
25 000~40 000	1	6	14	12
>40 000	0	1	9	11

来源:美国民意研究中心(National Opinion Research Center)的1996年综合社会调查(General Social Survey)。

我们通过表2.8加以说明。考虑一对研究对象,一个落在单元格(<15, VD),另一个落在单元格(15~25, LD)。这一对是相协的,因为第二个对象的收入和工作满意度都比第一个排得高。当单元格(<15, VD)中的对象与落在单元格(15~25, LD)中的三个对象的任何一个配对时都会形成相协对,因此这两个单元格提供了$1\times3=3$个相协对。当单元格(<15, VD)中的对象与任何一个在两个变量上都较其排序更高的对象($10+7+6+14+12+1+9+11$)配对时,就会形成相协对。同样地,当在单元格(<15, LD)中的三个对象与两个变量的排序都比其高的对象($10+7+14+12+9+11$)配对时也会构成相协对。

用C来表示相协对的总数,即有

$$C = 1(3+10+7+6+14+12+1+9+11)+3(10+7+14+12+9+11)+10(7+12+11)+2(6+14+12+1+9+11)+3(14+12+9+11)+10(12+11)+1(1+9+11)+6(9+11)+14(11)=1\,331。$$

类似地,相异对的总数是

$$D = 3(2+1+0)+10(2+3+1+6+0+1)+\cdots+12(0+1+9)=849。$$

在这个例子中,$C>D$,表明存在一种低收入伴随着低工作满意度而高收入伴随着高工作满意度的趋势。

考虑两个来自于联合概率分布$\{\pi_{ij}\}$的独立观测值。这两个观测值为相协对和相异对的概率分别为

$$\prod_c = 2\sum_i\sum_j \pi_{ij}\Big(\sum_{h>i}\sum_{k>j}\pi_{hk}\Big),\quad \prod_d = 2\sum_i\sum_j \pi_{ij}\Big(\sum_{h>i}\sum_{k<j}\pi_{hk}\Big)。$$

在这里,i 和 j 在括号内部的求和运算中保持不变,乘以 2 是因为有可能第一个观测值在单元格(i,j)而第二个观测值在单元格(h,k),也可能正好相反。几种关于定序变量的关联指标都与 $\prod_c - \prod_d$ 的差有关。

2.4.4 关联的定序度量指标:γ

给定在一个配对中两个变量都不相平的情况下,$\prod_c/(\prod_c + \prod_d)$ 是相协发生的概率,$\prod_d/(\prod_c + \prod_d)$ 是相异发生的概率。两个概率的差等于

$$\gamma = \frac{\prod_c - \prod_d}{\prod_c + \prod_d}, \tag{2.14}$$

被称为 γ(*gamma*,Goodman and Kruskal, 1954)。它的样本形式为 $\hat{\gamma} = (C - D)/(C + D)$。

与相关系数一样,γ 将两个变量视为对称的——不设定哪一个是结果变量。另外,γ 的取值范围也是 $-1 \leqslant \gamma \leqslant 1$。将一个变量的类别排序颠倒会引起 γ 的符号发生变化。尽管当 X 和 Y 之间存在完全线性的关系时它们之间的相关系数的绝对值才为 1,但是只要存在单调性就有 $|\gamma| = 1$。当 $\prod_d = 0$ 时,$\gamma = 1$;当 $\prod_c = 0$ 时,$\gamma = -1$。独立性意味着 $\gamma = 0$,反之,$\gamma = 0$ 并不保证变量间相互独立。例如,在一个 U 形的联合分布中就有可能出现 $\prod_c = \prod_d$,因而 $\gamma = 0$。

2.4.5 工作满意度例子的 γ

以表 2.8 为例,$C = 1\ 331$,$D = 849$。因此,

$$\hat{\gamma} = (1\ 331 - 849)/(1\ 331 + 849) = 0.221。$$

随着收入的增加,工作满意度仅仅存在着一个微弱的上升趋势。在非相平的配对中,相协对所占的比例比相异对高 0.221。

注 解

第 2.2 节:两个比例的比较

2.1 Breslow(1996)对个案-控制研究方法的发展过程进行了有趣的综述。

2.2 Edwards(1963)证明,对于 2×2 表格,有关发生比之比的函数是唯一的在进行行列互换或在行内、列内乘以某个常数后都不发生变动的统计量。Altham(1970)给出了关于 $I \times J$ 表格的相应结果。Yule(1912:587)曾经强调乘数恒定(multiplicative invariance)是关联度量指标的一个很好的特性,尤其是针对不同的边际类别采用不同抽样比例的情况。Goodman(2000)介绍了五种度量 2×2 表格的关联的方法,并给出了一个包含所有五种指标的一般性度量指标。

第 2.3 节:分层 2×2 表格中的偏关联

2.3 Paik(1985)提出用类似于图 2.2 的形式来描述三维表格。Friendly(2000)讨论了有关分类数据的图形表述。关于对辛普森悖论的更多讨论以及它在什么情况下发生,参见:Blyth(1972)、Davis(1989)、Dong(1998)、Samuels(1993)、Simpson(1951)。Good 和 Mittal(1991)将其扩展为合并悖论(*amalgamation paradox*),用来表示边际度量指标大于分表中相同指标的最大值或小于其最小值的情况。

第 2.4 节:扩展到 $I \times J$ 表格

2.4 对于连续变量，样本可以实现完全的排序（即不会出现相平的情况），因此 $C+D=\binom{n}{2}$ 以及 $\hat{\gamma}=(C-D)/\binom{n}{2}$。这被称为 Kendall 的 τ（*Kendall's tau*）。Agresti（1984，Chaps. 9 and 10）以及 Kruskal（1958）对关联的定序度量指标进行了综述。这些指标同样适用于当一个变量是定序的而另一个是二分变量的情况。当 Y 是定序变量，而 X 是 $I>2$ 的定类变量时，第 2.4 节所介绍的度量指标都没有什么实际意义。针对这种情况，定序模型方法（第 7.2 节）对 X 的每个类别都设定一个参数，通过这些参数可以比较在 X 的不同类别间定序结果变量的取值情况。

习题

应用部分

2.1 一篇《纽约时报》（*New York Times*，1999 年 2 月 17 日）的文章在介绍通过 PSA 血检诊断前列腺癌时指出："该检验对于每 4 个男性患者中会有 1 个无法检测出来（假阴性结果），并且有高达 2/3 的男性在检验中得到假阳性的结果。"令 $C(\bar{C})$ 表示患有（未患有）前列腺癌，并令 +（−）表示阳性（阴性）的结果。以下哪一个是正确的：$P(-|C)=\frac{1}{4}$ 还是 $P(C|-)=\frac{1}{4}$？$P(\bar{C}|+)=\frac{2}{3}$ 还是 $P(+|\bar{C})=\frac{2}{3}$？给出该检验的灵敏度和准确度。

2.2 一种诊断检验的灵敏度 = 准确度 = 0.80。求真实患病情况与诊断结果之间的发生比之比。

2.3 表 2.9 是基于佛罗里达州高速公路安全与机动车管理局所汇总的 1988 年发生的事故记录。指出哪一个是结果变量，求相应的比例之差、相对风险及发生比之比，并加以解释。为什么相对风险与发生比之比的值几乎相等？

表 2.9 习题 2.3 的数据

安全装置的使用	受伤情况	
	死 亡	未死亡
无	1 601	162 527
安全带	510	412 368

来源：佛罗里达高速公路安全与机动车管理局。

2.4 考虑《纽约时报》报道的以下两项研究结果。

a. 一项英国研究发现（1999 年 12 月 3 日），在罹患肺癌的吸烟者中，"女性患小细胞肺癌的可能性比男性高 1.7 倍"。在这里，1.7 是指发生比之比还是相对风险？

b. 国家癌症研究所进行的一项关于三苯氧胺与乳腺癌的研究发现（1998 年 4 月 7 日），服用三苯氧胺的妇女比服用安慰剂的妇女患扩散性乳腺癌的可能性低 45%。求下列相对风险：(i) 服用药物的与服用安慰剂的相比，(ii) 服用安慰剂的与服用药物的相比。

2.5 一项研究（E. G. Krug et al.，*Internat. J. Epidemiol.*，27：214-221，1998）发现，美国 1994 年每 100 000 人中死于枪击事件的人数是 14.24，加拿大是 4.31，澳大利亚是 2.65，德国是 1.24，以及英格兰和威尔士为 0.41。利用相对风险将美国与其他国家进行比较，并加以解释。

2.6 1994 年世界杯意大利和保加利亚的半决赛开赛前，一篇报纸文章写到："看好意大

利击败保加利亚的为10-11,而认为保加利亚进决赛的为10-3。"假设这意味着意大利赢球的发生比为$\frac{11}{10}$,而保加利亚赢球的发生比为$\frac{3}{10}$。求每个球队赢球的概率,并加以评论。

2.7 在美国,估计的35岁以上的妇女每年死于肺癌的概率对当前吸烟者来说是0.001 304,对非吸烟者是0.000 121(M. Pagano and K. Gauvreau, *Principles of Biostatistics*, Duxbury Press, Pacific Grove, CA. 1993: 134)。

a. 求比例之差与相对风险,并加以解释。就这些数据而言,哪个指标更有意义?为什么?

b. 求发生比之比并加以解释。说明为什么相对风险与发生比之比的值很相近。

2.8 在泰坦尼克号(*Titanic*)那次致命远航的成年乘客中,性别(女,男)与幸存(是,否)的发生比之比是11.4(数据可参见 R. J. M. Dawson, *J. Statist. Ed*. 3, 1995)。

a. "女性存活的概率是男性的11.4倍"这样的解释错在哪里?给出正确的解释。在什么情况下上述解释是大致正确的?

b. 女性幸存的发生比是2.9。计算分性别的幸存者比例。

2.9 在一篇关于美国犯罪问题的文章中,《新闻周刊》(*Newsweek*,1994年1月10日)引用美国联邦调查局在1992年的统计数据指出,在涉及黑人的凶杀案中94%的凶手为黑人,在涉及白人的凶杀案中83%的凶手为白人。令Y=受害人的种族,X=凶手的种族。这些统计数据所指的是哪一个条件分布,$Y|X$还是$X|Y$?如果要估计在给定凶手是白人的情况下受害人是白人的概率,你还需要哪些信息?求发生比之比并加以解释。

2.10 一项研究估计,在一定状况下,白人患者被推荐去做心脏导管插入术的概率是0.906,而黑人患者是0.847。

a. 这项研究的一次发布会声称,黑人被推荐做心脏导管插入术的发生比是白人的60%。说明他们是怎么得出的60%这个数(更精确地,57%)。

b. 美联社(Associated Press)的后续报道介绍了该研究,并指出:"医生们让黑人去做心脏导管插入术的可能性仅仅是白人的60%。"说明这一解释错在哪里。给出这种解释中正确的百分比(在向公众介绍结果时,使用相对风险要好于使用发生比之比。前者较容易理解也较少被误读。有关细节,参见 *New Engl. J. Med.* 341: 279-283, 1999)。

2.11 一项关于英国男性医师的20年队列研究(R. Doll and R. Peto, *British Med. J.* 2: 1525-1536, 1976)发现,吸烟者每年死于肺癌的比例是0.001 40,非吸烟者是0.000 10。吸烟者死于心脏病的比例是0.006 69,非吸烟者为0.004 13。

a. 使用比例之差、相对风险,以及发生比之比,分别描述吸烟与肺癌和吸烟与心脏病的关联,并加以解释。

b. 如果按照消除吸烟后死亡人数的下降来计算,哪一种死因与吸烟的关联更强?加以说明。

2.12 表2.10给出的是1973年秋季加州大学伯克利分校研究生院的申请情况。表中列出了六个最大的院系按照申请人性别划分的录取决定的分布。将这三个变量分别记为:A=是否被录取,G=性别,D=院系。求关于AG的样本条件发生比之比和边际发生比之比。解释结果,并说明为什么它们给出的有关AG的关联不相同。

表 2.10 习题 2.12 的数据

院 系	是否被录取			
	男 性		女 性	
	是	否	是	否
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317
小计	1 198	1 493	557	1 278

来源:数据取自:Freedman et al. (1978:14)。另见:P. Bickel et al. ,*Science* 187:398-403(1975)。

2.13 在"现实世界"中找出三个变量 X,Y 和 Z,使其满足在 X 和 Y 之间存在边际关联,但当控制了 Z 后 X 和 Y 条件独立。

2.14 根据美国 1987 年的谋杀率,美联社报道指出,一个新出生的婴儿最终被谋杀的概率对于非白人男性而言是 0.026 3,白人男性是 0.004 9,非白人女性是 0.007 2,白人女性是 0.002 3。

a. 求在给定性别的情况下种族与是否被谋杀之间的条件发生比之比,并对结果加以解释。这些变量存在同质关联吗?

b. 假设每个种族出生的新生儿的性别是一半一半。求种族与是否被谋杀之间的边际发生比之比。

2.15 南卡罗来纳州每个年龄的死亡率都比缅因州高,但是总的死亡率却比缅因州低。说明为什么会发生这种情况(数据参见:H. Wainer, *Chance* 12: 44, 1999)。

2.16 一项关于 1976—1991 年间肯塔基州死刑案件的研究(T. Keil and G. Vito, *Amer. J. Criminal Justice* 20: 17-36, 1995)发现,在白人杀死白人的 391 件案子中被告人被判死刑的为 8%,在 102 件黑人杀死黑人的案子中相应比例为 2%,在 57 件黑人杀死白人的案子中被判死刑的占 12%,在 18 件白人杀死黑人的案子中为 0%。构建一个三维列联表,求被告人种族与死刑判决结果之间的条件发生比之比,解释这些关联,并分析是否存在辛普森悖论的情形。说明为什么边际关联与条件关联之间差别很大。

2.17 利用成年女性罹患扁平细胞癌(是,否)与吸烟行为(吸烟,不吸烟)的数据估计的发生比之比:在每日吸烟水平 s 为 $0<s<20$ 根时等于 11.7,$s\geqslant 20$ 时为 26.1(R. C. Brownson et al. , *Epidemiology* 3: 61-64, 1992),证明关于癌症(是,否)与吸烟水平($s\geqslant 20$,$0<s<20$)的发生比之比的估计值等于 2.2。

2.18 表 2.11 给出的是一项在几家英国医院进行的关于是否患肺癌与吸烟行为的回顾性研究。该表按照在发病前 10 年间每天平均吸烟数将男性肺癌患者与患其他病的控制组病人进行了对比。

a. 求在每个吸烟水平上肺癌的样本发生比,以及非吸烟者分别与另外五个吸烟水平相比的发生比之比。当吸烟水平上升时,是否存在某种趋势?对结果加以解释。

b. 如果肺癌的对数发生比与吸烟水平线性相关,即第 i 行的对数发生比满足 $\log(\text{发生比}_i)=\alpha+\beta i$,证明局部发生比之比都相等。

c. 利用这些数据,你能不能估计在每个吸烟水平上肺癌发生的概率?在(a)中估

计的发生比之比有意义吗？请解释。

d. 证明针对不同吸烟水平而言，各疾病种类的分布是*随机排序的*(*stochastically ordered*)(参见习题 2.34 以及第 7.3.4 节)。加以解释。

表 2.11 习题 2.18 的数据

日均吸烟量	疾病种类	
	肺癌患者	其他患者
0	7	61
<5	55	129
5 ~ 14	489	570
15 ~ 24	475	431
25 ~ 49	293	154
≥50	38	12

来源：授权重印自：R. Doll and A. B. Hill, *British Med. J.* 2:1271-1286(1952)。

2.19 表 2.12 给出了亚利桑那州 91 对已婚夫妇关于性享受频繁程度的回答。给出一个关于妻子的回答与丈夫的回答之间的关联度量指标，并加以解释。

表 2.12 习题 2.19 的数据

丈夫的评分	妻子关于性享受的评分			
	从未或偶尔	比较频繁	非常频繁	几乎总是
从未或偶尔	7	7	2	3
比较频繁	2	8	3	7
非常频繁	1	5	4	9
几乎总是	2	8	9	14

来源：授权重印自：Hout et al. (1987)。

2.20 表 2.13 来自于一项关于佛罗里达州死刑判决的早期研究。分析这些数据并证明该数据中存在辛普森悖论。

表 2.13 习题 2.20 的数据

受害人种族	被告人种族	死刑判决	
		是	否
白人	白人	19	132
	黑人	11	52
黑人	白人	0	9
	黑人	6	97

来源：授权重印自：M. L. Radelet, *Amer. Sociol. Rev.* 46:918-927(1981)。

理论与方法

2.21 在一项关于某种疾病的诊断检测中，π_1 表示在研究对象患病的情况下诊断为阳性的概率，π_2 表示在研究对象没有患病的情况下诊断为阳性的概率。令 ρ 表示研究对象确实患有此病的概率。

a. 给定诊断结果为阳性，证明对象确实患有此病的概率为

$$\pi_1\rho/[\pi_1\rho + \pi_2(1-\rho)]。$$

b. 假定一项关于 HIV 阳性的诊断检查中的灵敏度和准确度都等于 0.95，且 $\rho =$ 0.005，求给定诊断结果为阳性的情况下，研究对象确实是 HIV 阳性的概率。为了更好地理解这一问题，求诊断结果与实际患病状态的联合概率，并讨论它们

的相对大小。

2.22 利用图示法来描述两组的二项分布参数,分别用水平坐标轴表示 π_1,垂直坐标轴表示 π_2。画出满足以下条件的一个 2×2 表格的点的轨迹:(a)相对风险等于0.5,(b)发生比之比等于0.5,(c)比例之差等于 -0.5。

2.23 令 D 表示患有某种疾病,E 表示接触过某种风险因素。*可解释风险*(*attributable risk*, AR)是指病例中可归因于该风险因素的比例(参见:Benichou, 1998)。

a. 令 $P(\bar{E}) = 1 - P(E)$。解释为什么

$$\mathrm{AR} = [P(D) - P(D \mid \bar{E})]/P(D)。$$

b. 证明 AR 与相对风险 RR 存在以下关系:

$$\mathrm{AR} = [P(E)(\mathrm{RR}-1)]/[1 + P(E)(\mathrm{RR}-1)]。$$

2.24 对于一个计数为 $\{n_{ij}\}$ 的 2×2 表格,证明在以下情况下发生比之比不变:(a)行和列互换,(b)在行内或列内将单元格计数乘以一个 $c\neq0$ 的常数。证明比例之差和相对风险不具备这些特性。

2.25 对于给定的 π_1 和 π_2,当二者独立时,发生比之比和相对风险等于1.0,证明在任何情况下相对风险与1.0的距离不会超过发生比之比与1.0的距离。

2.26 解释为什么对于三个事件 E_1,E_2,E_3 以及它们的补集,即使 $P(E_1 \mid E_2E_3) < P(E_1 \mid \bar{E}_2E_3)$ 和 $P(E_1 \mid E_2\bar{E}_3) < P(E_1 \mid \bar{E}_2\bar{E}_3)$ 都成立时,仍然有可能存在 $P(E_1 \mid E_2) > P(E_1 \mid \bar{E}_2)$(提示:应用有关三维表格的辛普森悖论)。

2.27 令 $\pi_{ij|k} = P(X=i, Y=j \mid Z=k)$,解释为什么 XY 条件独立等同于对于所有 i,j,k,$\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}$。

2.28 对于一个 $2\times2\times2$ 表格,通过显示所有 XY 的条件发生比之比相等等价于所有 YZ 的条件发生比之比相等,以表明同质关联具有对称性特征。

2.29 史密斯和琼斯都是棒球运动员。在 K 年中,史密斯每一年的击球率都比琼斯高。有没有可能将 K 年的数据合并后,琼斯的击球率比史密斯高?请加以解释,并举例说明。

2.30 当 X 和 Y 在 Z 的每个取值上都条件相依但二者边际独立时,Z 被称为一个*抑制变量*(*suppressor variable*)。指定一个 $2\times2\times2$ 表格的联合分布证明,在以下条件下有可能发生上述情形:(a)存在同质关联,(b)分表中的关联方向相反。

2.31 证明公式2.11中的 $\{\alpha_{ij}\}$ 决定了:(a)所有由任意两行与任意两列所形成的 $\binom{I}{2}\binom{J}{2}$ 个发生比之比,(b)公式2.10中的所有 $\{\theta_{ij}\}$,反之亦然。

2.32 参见习题2.31。证明当所有行和列的概率都为正时,独立性等价于所有 $\{\alpha_{ij}=1\}$。

2.33 对于 $I\times J$ 的列联表,说明为什么当 $(I-1)(J-1)$ 个差 $\pi_{j|i} - \pi_{j|I} = 0$ 时,变量之间相互独立,其中 $i=1,\cdots,I-1, j=1,\cdots,J-1$。

2.34 对于一个具有定序结果变量的 $2\times J$ 表格,令 $F_{j|i} = \pi_{1|i} + \cdots + \pi_{j|i}$,当对所有 $j=1,\cdots,J$ 都存在 $F_{j|2} \leqslant F_{j|1}$ 时,则称第2行中的条件分布*随机高于*(*stochastically higher*)第1行中的条件分布。考虑*累积发生比之比*(*cumulative odds ratios*)

$$\theta_j = \frac{F_{j|1}/(1-F_{j|1})}{F_{j|2}/(1-F_{j|2})}, \qquad j = 1,\cdots,J-1。$$

a. 证明如果对所有 j 都有 $\log\theta_j \geqslant 0$,则有第2行随机高于第1行。说明为什么在这种情况下第2行的观测值比第1行更可能落在定序尺度的较高一端。

b. 如果所有的局部对数发生比之比均非负,对于 $1\leqslant j\leqslant J-1$,$\log\theta_j \geqslant 0$(Lehmann,

1966)。通过举反例证明其反命题不成立。

2.35 假设$\{Y_{ij}\}$是均值为$\{\mu_{ij}\}$的独立的泊松变量。证明对于所有i和j,在$\{Y_{i+}=n_i\}$的条件下,$P(Y_{ij}=\mu_{ij})$在各行内满足独立的多项分布抽样(即公式2.2对所有i的乘积)。

2.36 对于2×2表格,Yule(1900,1912)引入了

$$Q=\frac{\pi_{11}\pi_{22}-\pi_{12}\pi_{21}}{\pi_{11}\pi_{22}+\pi_{12}\pi_{21}},$$

为了纪念比利时统计学家Quetelet,他将此记为Q。现在这个指标被称为Yule的Q (*Yule's Q*)。

a. 证明对于2×2表格,Goodman 和 Kruskal 的$\gamma=Q$。

b. 证明Q的取值范围为-1到1。

c. 给出$Q=-1$或$Q=1$时的条件。

d. 证明Q与发生比之比的关系为$Q=(\theta-1)/(\theta+1)$,它是一个将θ从$[0,\infty]$尺度到$[-1,+1]$尺度的单调转换。

2.37 当X和Y是计数为$\{n_{ij}\}$的定序变量时:

a. 说明为什么$\binom{n}{2}$个观测值的配对可分割为$C+D+T_X+T_Y-T_{XY}$,其中$T_X=\sum n_{i+}(n_{i+}-1)/2$为关于$X$相平的配对数,$T_Y$是关于$Y$相平的配对数,$T_{XY}$是关于$X$和$Y$都相平的配对数。

b. 对于每一对排序的观测值(X_a,Y_a)和(X_b,Y_b),令$X_{ab}=$符号(X_a-X_b),$Y_{ab}=$符号(Y_a-Y_b)。证明有关$n(n-1)$个独立的(X_{ab},Y_{ab})配对的样本相关系数为

$$\tau_b=\frac{C-D}{\left\{\left[\binom{n}{2}-T_X\right]\left[\binom{n}{2}-T_Y\right]\right\}^{1/2}}。$$

这一定序度量指标被称为Kendall的τ_b(*Kendall's tau-b*,*Kendall*, 1945)。与γ相比,τ_b对结果变量的类别划分较不敏感。

c. 令$d=(C-D)/\left[\binom{n}{2}-T_X\right]$,解释为什么$d$是在$X$上不相平的配对中相协对与相异对的比例之差(Somers, 1962)(对于2×2表格,d等于比例之差,τ_b等于X和Y的相关系数)。

2.38 Goodman 和 Kruskal(1954)提出了一个度量定类变量之间关联的指标(τ),它与变异程度的度量指标$V(Y)$有关,

$$V(Y)=\sum\pi_{+j}(1-\pi_{+j})=1-\sum\pi_{+j}^2。$$

a. 证明$V(Y)$是两个关于Y的独立观测值落入不同类别的概率(称为*基尼集中指数*,*Gini concentration index*)。证明当存在j使得$\pi_{+j}=1$时,$V(Y)=0$;当对于所有j都有$\pi_{+j}=1/J$时,$V(Y)$取其最大值$(J-1)/J$。

b. 对于变异消减比例,证明$E[V(Y|X)]=1-\sum_i\sum_j\pi_{ij}^2/\pi_{i+}$(推导所得的度量指标(式2.12)被称为*集中系数*(*concentration coefficient*)。与U相似,$\tau=0$等价于X和Y相互独立。Haberman(1982)给出了一般化的集中系数和不确定系数)。

2.39 对于度量定类变量关联的指标λ(Goodman and Kruskal, 1954),有$V(Y)=1-\max\{\pi_{+j}\}$以及$V(Y|i)=1-\max_j\{\pi_{j|i}\}$,解释$\lambda$是用一个最可能发生的结果类别来进行预测时对应的预测错误消减比例。证明独立性意味着$\lambda=0$,但其反命题不成立。

3 列联表的统计推断

在这一章,我们介绍关于列联表的统计推断方法。本章介绍的许多方法对于后面章节有关非列联表形式的分类数据的分析同样重要。这些方法通常假定泊松分布抽样、多项分布抽样或独立的二项分布抽样。

在第3.1节,我们介绍包括发生比之比的有关 2×2 表格关联度量指标的置信区间。第3.2节的内容是对两个分类变量之间独立性假设的卡方检验。与其他显著性检验一样,它的实际价值有限。第3.3节介绍如何利用残差或卡方的可分割性对检验做进一步的分析。第3.4节给出了针对定序变量的更有效的统计推断方法。这4节介绍的方法仅适用于大样本的情况。在第3.5和3.6节中,我们介绍有关小样本的统计推断方法。

3.1 关联参数的置信区间

关联参数估计值的精确度可以用相应样本分布的标准误来表示。在这节中,我们给出大样本情况下的标准误和置信区间。

3.1.1 发生比之比的区间估计

对于 2×2 表格,样本发生比之比为 $\hat{\theta}=n_{11}n_{22}/n_{12}n_{21}$。当其中任何一个 $n_{ij}=0$ 时,样本发生比之比等于0或∞;如果某一行或某一列的两个单元格都为零,样本发生比之比就无法定义。由于出现这些情况的概率不为0,因而 $\hat{\theta}$ 和 $\log\hat{\theta}$ 的期望值和方差都不存在(事实上,这一点对于后面章节中将要介绍的有关模型参数的最大似然估计值同样适用)。就误差和均方差(mean-squared error)而言,Gart 和 Zweiful(1967)以及 Haldane(1956)都表明,修正后的估计值

$$\tilde{\theta}=\frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)}$$

以及相应的 $\log\tilde{\theta}$ 表现良好(习题14.4)。

估计值 $\hat{\theta}$ 和 $\tilde{\theta}$ 具有同样的关于 θ 的渐近正态分布。但是,除非样本规模 n 非常大,它们的分布是高度偏斜的。例如,当 $\theta=1$ 时,$\hat{\theta}$ 不能取比 θ 小很多的值(因为 $\hat{\theta}\geq0$),但是它可以取比 θ 大很多的值,而且这种情况发生的概率往往无法忽略。它的对数转换形式具有可加的而非可积的结构,并以更快的速度收敛于正态分布。关于 $\log\hat{\theta}$ 的估计标准误为

$$\hat{\sigma}(\log\hat{\theta}) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{1/2}。\tag{3.1}$$

有关推导参见第 3.1.7 节。

根据 $\log\hat{\theta}$ 在大样本条件下服从正态分布,

$$\log\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}(\log\hat{\theta})\tag{3.2}$$

就是 $\log\theta$ 的沃尔德置信区间(Wald confidence interval)。对区间的两个端点分别求幂(即反对数)便得出关于 θ 的置信区间。这个区间是由 Woolf(1955)提出的,其表现非常好,结果往往略为保守(即它实际涵盖的概率大于其名义水平)。

当 $\hat{\theta}=0$ 或 ∞ 时,Woolf 区间不存在。当 $\hat{\theta}=0$ 时,应当用 0 作为区间的下限;当 $\hat{\theta}=\infty$ 时,用 ∞ 作为区间的上限。也可以通过对 Woolf 公式进行一定的调整来计算区间的上下限,如 Gart(1966)区间使用 $\{n_{ij}+0.5\}$ 而不是 $\{n_{ij}\}$ 来计算估计值及其标准误。一种更为通行的办法是,通过对 θ 的计分检验(Cornfield, 1956)或似然比检验求逆来构建置信区间,我们将在第 3.1.8 节中对此方法加以讨论。

3.1.2 例子:阿司匹林与心肌梗塞

我们利用表 3.1 来举例说明发生比之比的统计推断,与第 2.2.5 节所介绍的例子相似,该表的数据取自瑞典一项有关服用阿司匹林与心肌梗塞关系的研究。该研究将1 360 名心脏病患者随机分为两组:一组服用阿司匹林(每天一小片),另一组服用安慰剂。表 3.1 给出了随后跟踪的 3 年中两组分别死于心肌梗塞的人数。

样本的发生比之比 $\hat{\theta}=1.56$,与 $\tilde{\theta}=1.55$ 非常相近,这是因为所有单元格计数都特别小。由公式 3.1 可得,$\log\hat{\theta}=0.445$ 的标准误为 $\hat{\sigma}(\log\hat{\theta})=0.307$。

表 3.1 瑞典关于服用阿司匹林与心肌梗塞关系的研究

	心肌梗塞		
	是	否	小计
安慰剂	28	656	684
阿司匹林	18	658	676

来源:基于 Lancet (Lancet338:1345-1349,1991)所给出的结果。

在这个样本所代表的总体中,关于 $\log\theta$ 的 95% 的置信区间为 $0.445 \pm 1.96(0.307)$,即$(-0.157, 1.047)$。θ 所对应的区间为$[\exp(-0.157), \exp(1.047)]$,即$(0.85, 2.85)$。由此可见,该结果对真实发生比之比的估计相当不精确。

由于 θ 的置信区间包含 1.0,有可能服用阿司匹林和安慰剂的患者,死于心肌梗塞的真实发生比是一样的。因为出现心肌梗塞死亡的案例相对较少,如果服用阿司匹林确实有效,但是相应的发生比之比不是很大,那么可能需要一个足够大的样本才能显现这种作用(习题 3.21)。

3.1.3 比例之差的区间估计

比例之差和相对风险用来比较结果变量在两组间的条件分布。对于这些指标,我们将样本视为独立的二项分布样本。也即,对于第 i 组,y_i 服从一个样本规模为 n_i、"成功"结果的概率为 π_i 的二项分布。

样本比例 $\hat{\pi}_i = y_i/n_i$ 的期望值为 π_i,方差为 $\pi_i(1-\pi_i)/n_i$。由于 $\hat{\pi}_1$ 和 $\hat{\pi}_2$ 相互独立,

二者之差的期望值为

$$E(\hat{\pi}_1 - \hat{\pi}_2) = \pi_1 - \pi_2,$$

标准误为

$$\sigma(\hat{\pi}_1 - \hat{\pi}_2) = \left[\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right]^{1/2}。\tag{3.3}$$

将公式 3.3 中的 π_i 替换为 $\hat{\pi}_i$，就得到了 $\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2)$ 的估计值。那么

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2) \tag{3.4}$$

是关于 $\pi_1 - \pi_2$ 的沃尔德置信区间。与式 1.13 所表示的有关单个比例的沃尔德区间相似，通常它所涵盖的真实概率小于名义的置信水平，尤其是当 π_1 和 π_2 的值接近于 0 或 1 的时候。在第 3.1.8 节、注解 3.2，以及习题 3.23 中，我们给出了相对复杂但效果更好的方法。

3.1.4 相对风险的区间估计

样本的相对风险为 $r = \hat{\pi}_1/\hat{\pi}_2$。与发生比之比相类似，相对风险在取对数后会更快地收敛于正态分布。$\log r$ 的渐近标准误（asymptotic standard error）是

$$\sigma(\log r) = \left(\frac{1-\pi_1}{\pi_1 n_1} + \frac{1-\pi_2}{\pi_2 n_2}\right)^{1/2}。\tag{3.5}$$

可以对 $\log r \pm z_{\alpha/2}\hat{\sigma}(\log r)$ 的端点分别求幂，得出相对风险的沃尔德区间。该区间表现良好，但相对保守。在第 3.1.8 节中，我们将讨论另一种可行的方法。

对于表 3.1，服用安慰剂的对象死于心肌梗塞的样本比例为 0.040 9，服用阿司匹林的相应比例为 0.026 6。样本的相对风险等于 0.040 9/0.026 6 = 1.54。对数相对风险的 95% 的置信区间为 log(1.54) ± 1.96(0.297)，相应地，相对风险的置信区间为(0.86, 2.75)。由此推论，服用安慰剂患者的死亡率是服用阿司匹林患者的 0.86 到 2.75 倍。$\pi_1 - \pi_2$ 的 95% 的沃尔德置信区间是 0.014 ± 1.96(0.009 8)，即(−0.005, 0.033)。无论按照哪一个指标，服用阿司匹林既可能带来巨大的公共卫生回报，也有可能没有效果或存在微小的负面效应。而第 2.2.5 节介绍的更大规模的研究结果显示，服用阿司匹林确实是有效的。

3.1.5 应用 δ 方法推导标准误*

在大样本情况下，有一种简单而有效的推导标准误的方法。令 T_n 表示一个渐近服从参数为 θ 的正态分布的统计量，下标 n 表示该统计量的分布取决于样本规模。假定估计值是 T_n 的一个函数 $g(T_n)$，那么，在非极端的情况（mild conditions）下，$g(T_n)$ 本身也服从大样本正态分布，其标准误取决于当 t 的值接近于 θ 时 $g(t)$ 的变化速度。

具体地说，对于大样本 n，假定 T_n 服从参数为 θ 的正态分布且标准误为 $\sigma/\sqrt{n}$。也即，当 $n\to\infty$ 时，$\sqrt{n}(T_n - \theta)$ 的累积分布函数收敛于均值为 0、方差为 σ^2 的正态随机变量的累积分布函数。这种极限特性展示了一种*依分布收敛*（*convergence in distribution*）的情况，用符号表示为：

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)。$$

令 g 表示一个在 θ 处至少二次可导的函数，利用 $g(t)$ 在 $t = \theta$ 处的泰勒级数展开，对于大样本 n 有（详见第 14.1.2 节）：

$$\sqrt{n}[g(T_n) - g(\theta)] \approx \sqrt{n}(T_n - \theta)g'(\theta),$$

其中 $g'(\theta) = \partial g/\partial t$ 在 $t=\theta$ 处的取值。已知对于任意的变量 $Y \sim N(0,\sigma^2)$，则有 $cY \sim N(0,c^2\sigma^2)$。因而，

$$\sqrt{n}\left[g(T_n) - g(\theta)\right] \xrightarrow{d} N\left(0,\left[g'(\theta)\right]^2\sigma^2\right)。\tag{3.6}$$

换句话说，$g(T_n)$近似服从参数为 $g(\theta)$的正态分布，其方差为$[g'(\theta)]^2\sigma^2/n$。

图 3.1 展示了这一结果。在 θ 附近，$g(t)$大致呈斜率为 $g'(\theta)$的直线。此时，$g(T_n)$近似服从正态分布，这是因为正态随机变量的线性转换仍服从正态分布。$g(T_n)$值相对于 $g(\theta)$的离散度，大约为 T_n 值相对于 θ 的离散度的$|g'(\theta)|$倍。如果 g 在 θ 处的斜率是 $\frac{1}{2}$，那么 g 所对应的 $g(T_n)$的值域仅相当于 T_n 值域的一半左右。

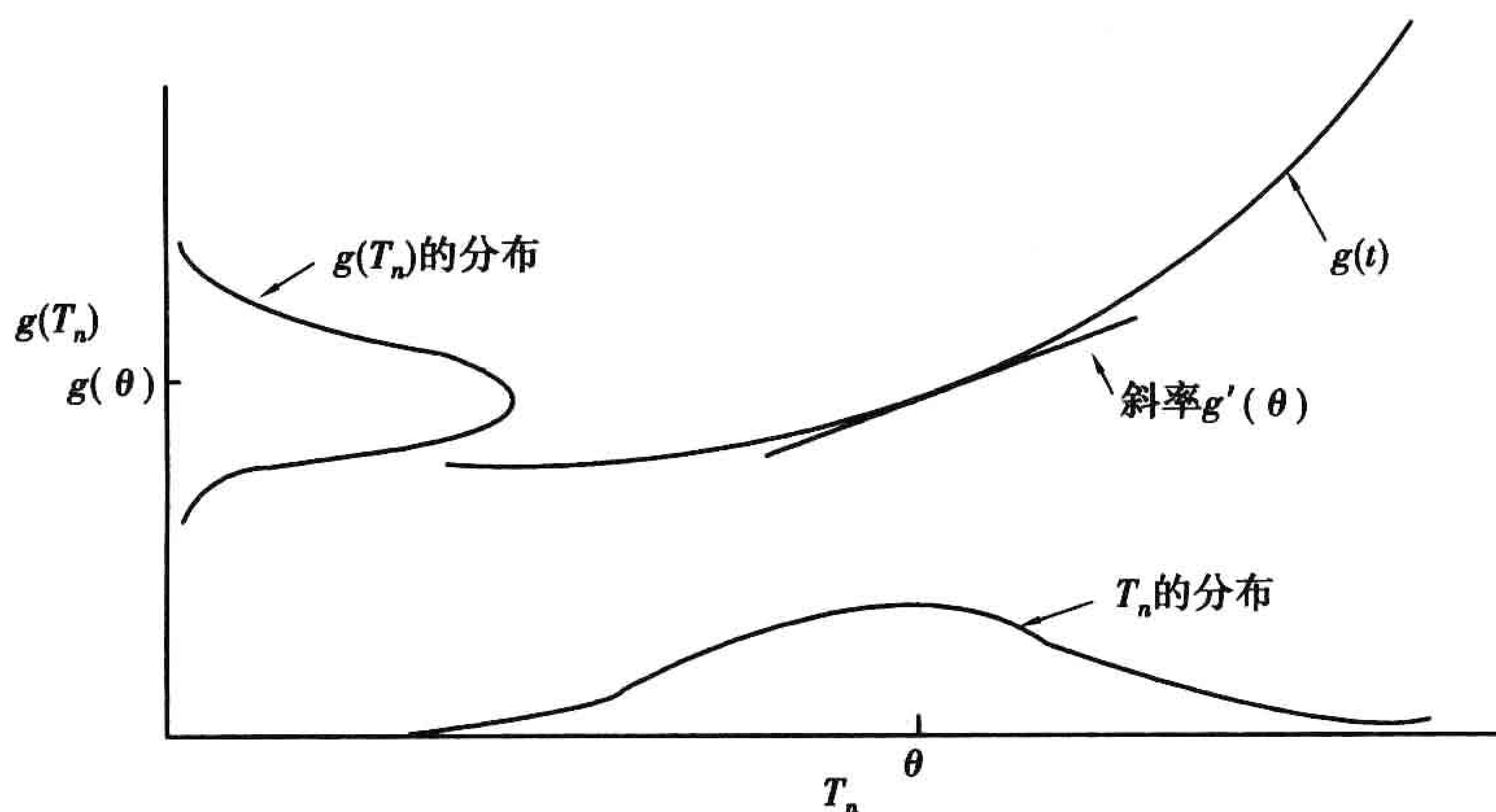


图 3.1 关于 δ 方法的描述

公式 3.6 的结果被称为 δ 方法（*delta method*）。由于 $g'(\theta)$和 $\sigma^2=\sigma^2(\theta)$通常取决于未知参数 θ，因此渐近方差是未知的。置信区间和统计检验中用 T_n 来替代 θ，并利用 $\sqrt{n}[g(T_n)-g(\theta)]/|g'(T_n)|\sigma(T_n)$渐近服从标准正态分布的特性。例如，

$$g(T_n) \pm 1.96|g'(T_n)|\sigma(T_n)/\sqrt{n}$$

是大样本情况下关于 $g(\theta)$的 95% 的沃尔德置信区间。

3.1.6 关于样本 Logit 的 δ 方法*

令在 n 次试验中成功 y 次的二项分布参数 π 的最大似然估计 $T_n=\hat{\pi}=y/n$，现在我们通过 T_n 函数来演示 δ 方法。由于 $E(Y)=n\pi$，$\text{var}(Y)=n\pi(1-\pi)$，所以有 $E(\hat{\pi})=\pi$ 以及 $\text{var}(\hat{\pi})=\pi(1-\pi)/n$。根据中心极限定理，$\hat{\pi}$ 服从大样本正态分布。关于 $\hat{\pi}$ 的函数多数也服从大样本正态分布。

$\hat{\pi}$ 的对数发生比函数

$$g(\hat{\pi}) = \log\left[\hat{\pi}/(1-\hat{\pi})\right],$$

被称为样本 Logit。对该函数的导数在 π 处求值，即等于 $1/\pi(1-\pi)$。按照 δ 方法，样本 Logit 的渐近方差为 $\pi(1-\pi)/n$（即 $\hat{\pi}$ 的方差）乘以$[1/\pi(1-\pi)]$的平方，也即，

$$\sqrt{n}\left(\log\frac{\hat{\pi}}{1-\hat{\pi}} - \log\frac{\pi}{1-\pi}\right) \xrightarrow{d} N\left(0,\frac{1}{\pi(1-\pi)}\right)。$$

由 $\hat{\pi}$ 渐近服从正态分布推导得出 $\log[\hat{\pi}/(1-\hat{\pi})]$也渐近服从正态分布。

渐近方差是指在大样本情况下与真实分布相近似的正态分布的方差。它不是对真实分布方差本身的近似。当 $0<\pi<1$ 时，样本 Logit 的渐近方差 $[n\pi(1-\pi)]^{-1}$ 是有解的。相反，真实的方差并不存在：因为出现 $\hat{\pi}=0$ 或 1 的概率为正，Logit 以不为零的概率等于 $-\infty$ 或 ∞。随着 n 的增加，Logit 等于无穷大的概率迅速收敛于零。在大样本的情况下，样本 Logit 的分布大致服从均值为 $\log[\pi/(1-\pi)]$、标准差为 $[n\pi(1-\pi)]^{-1/2}$ 的正态分布。因此，对于 Logit，渐近方差要比真实的方差更有意义。附带提一下，与此有关的是，重抽样自举法（bootstrap）无法用于求解许多离散度量指标的近似方差，因为它所近似的是真实的方差，而不是更有意义的渐近方差。

3.1.7 关于对数发生比之比的 δ 方法*

对数发生比之比和对数相对风险的标准误可以通过多参数形式的 δ 方法来计算。假设 $\{n_i, i=1,\cdots,c\}$ 服从参数为 $(n,\{\pi_i\})$ 的多项分布。样本比例 $\hat{\pi}_i=n_i/n$ 具有如下的均值和方差：

$$E(\hat{\pi}_i)=\pi_i \qquad \text{以及} \qquad \mathrm{var}(\hat{\pi}_i)=\pi_i(1-\pi_i)/n。 \tag{3.7}$$

当 $i\neq j$ 时，$\hat{\pi}_i$ 和 $\hat{\pi}_j$ 的协方差为（详见第 14.1.4 节）：

$$\mathrm{cov}(\hat{\pi}_i,\hat{\pi}_j)=-\pi_i\pi_j/n。 \tag{3.8}$$

样本比例 $(\hat{\pi}_1,\hat{\pi}_2,\cdots,\pi_{c-1})$ 服从大样本的多元正态分布。对于样本比例的函数，δ 方法意味着以下结果（有关证明参见第 14.1.4 节）：

令 $g(\boldsymbol{\pi})$ 表示关于 $\{\pi_i\}$ 的可导函数，$g(\hat{\boldsymbol{\pi}})$ 为该函数服从多项分布的样本值。令

$$\phi_i=\frac{\partial g(\boldsymbol{\pi})}{\partial \pi_i}, \quad i=1,\cdots,c。$$

那么，随着 $n\to\infty$，$\sqrt{n}[g(\hat{\boldsymbol{\pi}})-g(\boldsymbol{\pi})]/\sigma$ 的分布收敛于标准正态分布，其中

$$\sigma^2=\sum\pi_i\phi_i^2-\left(\sum\pi_i\phi_i\right)^2。 \tag{3.9}$$

渐近方差的大小取决于 $\{\pi_i\}$ 以及该函数对 $\{\pi_i\}$ 的偏导数。在应用中，将公式 3.9 中的 $\{\pi_i\}$ 和 $\{\phi_i\}$ 代入它们的样本值，便可得到关于 σ^2 的最大似然估计 $\hat{\sigma}^2$。在此基础上求得 $\hat{\sigma}/\sqrt{n}$，即 $g(\hat{\boldsymbol{\pi}})$ 的标准误的估计值。关于 $g(\boldsymbol{\pi})$ 的大样本沃尔德置信区间为

$$g(\hat{\boldsymbol{\pi}})\pm z_{\alpha/2}\hat{\sigma}/\sqrt{n}。$$

用 $\hat{\sigma}$ 替代公式 3.9 中的 σ，其极限分布仍然服从标准正态分布，但收敛的速度较慢。在大样本分布中，这种等价性是基于以下原因：按照大数定理的弱定理，样本比例依概率收敛于 $\{\pi_i\}$。由于 $\hat{\sigma}$ 是样本比例的连续函数，它依概率收敛于 σ，因而，$\sigma/\hat{\sigma}$ 依概率收敛于 1，故有

$$\sqrt{n}\,\frac{g(\hat{\boldsymbol{\pi}})-g(\boldsymbol{\pi})}{\hat{\sigma}}=\sqrt{n}\,\frac{g(\hat{\boldsymbol{\pi}})-g(\boldsymbol{\pi})}{\sigma}\,\frac{\sigma}{\hat{\sigma}}。$$

按照公式 3.9，等式右边的第一项依分布收敛于标准正态分布，第二项依概率收敛于 1。因此，二者乘积的极限也服从标准正态分布。

接下来，我们将 δ 方法应用于对数发生比之比，令 $g(\boldsymbol{\pi})=\log\theta=\log\pi_{11}+\log\pi_{22}-\log\pi_{12}-\log\pi_{21}$。由于

$$\phi_{11}=\partial(\log\theta)/\partial\pi_{11}=1/\pi_{11},$$
$$\phi_{12}=-1/\pi_{12}, \quad \phi_{21}=-1/\pi_{21}, \quad \phi_{22}=1/\pi_{22},$$

$\sum_i \sum_j \pi_{ij}\phi_{ij} = 0$ 且 $\sigma^2 = \sum_i \sum_j \pi_{ij}\phi_{ij}^2 = \sum_i \sum_j (1/\pi_{ij})$。关于多项分布样本 $\{n_{ij}\}$ 的 $\log \hat{\theta}$ 的渐近标准误为

$$\sigma(\log \hat{\theta}) = \sigma/\sqrt{n} = \left(\sum_i \sum_j 1/n\pi_{ij}\right)^{1/2}。$$

由于 $n\hat{\pi}_{ij} = n_{ij}$，标准误的估计值同公式 3.1。

δ 方法也可以直接应用于 θ，以计算 $\hat{\sigma}(\hat{\theta})$ 以及沃尔德置信区间 $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}(\hat{\theta})$。但我们并不推荐这种做法，因为与 $\log \hat{\theta}$ 相比，$\hat{\theta}$ 向正态分布收敛的速度更慢，这个区间可能包含负值，而且它不能给出与使用 $1/\hat{\theta}$ 及其标准误的沃尔德区间等价的结果。

3.1.8 计分和剖面似然置信区间*

通过 δ 方法得出的标准误对应的是沃尔德置信区间。然而，在小样本或中等样本的情况下，基于对沃尔德检验求逆所得的区间有时候无法让人满意。这种情况下，可以考虑使用似然比或计分检验反向推导置信区间。尽管这些计算过程更加复杂，但其结果往往更好。

我们首先介绍关于比例之差的计分法。针对 $H_0: \pi_1 - \pi_2 = \Delta$ 的计分检验（Mee, 1984; Miettinen and Nurminen, 1985）具有检验统计量

$$z(\Delta) = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \Delta}{\sqrt{\hat{\pi}_1(\Delta)[1 - \hat{\pi}_1(\Delta)]/n_1 + \hat{\pi}_2(\Delta)[1 - \hat{\pi}_2(\Delta)]/n_2}},$$

其中 $\hat{\pi}_i(\Delta)$ 表示满足限定条件 $\pi_1 - \pi_2 = \Delta$ 的 π_i 的最大似然估计。也即，$\hat{\pi}_1(\Delta)$ 和 $\hat{\pi}_2(\Delta)$ 是满足 $\pi_1 - \pi_2 = \Delta$ 并使两个二项分布概率密度函数的乘积最大化的 π_1 和 π_2 的值。这些值不能通过封闭形式的解来表达，而需要用数值法（numerical methods）来决定。计分置信区间就是满足 $|z(\Delta)| < z_{\alpha/2}$ 的一组 Δ。这样的区间需要通过迭代法来计算（Nurminen, 1986）。

同样，使用计分法所得到的有关相对风险的置信区间也优于沃尔德区间（Bedrick, 1987; Gart and Nam, 1988; Koopman, 1984; Miettinen and Nurminen, 1985; Nurminen, 1986）。Corfield（1956）、Miettinen 和 Nurminen（1985）给出了关于发生比之比的计分区间。我们倾向于不对这些区间进行连续性或有限样本校正，因为那样做会导致结果太保守。计分区间在运算上比沃尔德区间更复杂，但对于现代运算能力来说，这不会构成使用计分区间的障碍，况且其背后的原理非常简单明了。不过，目前常用的软件还不提供这些区间。

接下来，我们通过发生比之比来展示基于似然比检验的置信区间。2×2 表格的多项分布似然函数是关于 $\{\pi_{11}, \pi_{12}, \pi_{21}\}$ 的函数，它也可以由 $\{\theta, \pi_{1+}, \pi_{+1}\}$ 来等价表示（参见第 2.4.1 节）。这样，我们可以通过对似然比检验 $H_0: \theta = \theta_0$ 的反向推导来判断 θ_0 是否属于该置信区间。这里存在两个冗余参数（*nuisance parameters*），在零假设下最大化似然函数的初始最大似然估计值 $\hat{\pi}_{1+}(\theta_0)$ 和 $\hat{\pi}_{+1}(\theta_0)$ 随着 θ_0 的变动而变动。

剖面对数似然函数（*profile log-likelihood function*）可表示为 $L(\theta_0, \hat{\pi}_{1+}(\theta_0), \hat{\pi}_{+1}(\theta_0))$，它是关于 θ_0 的函数。对于每个 θ_0，该函数给出在满足限定条件 $\theta = \theta_0$ 下的普通对数似然函数的最大值。对该函数在 $\theta_0 = \hat{\theta}$ 处求值，就得到了最大化的对数似然值 $L(\hat{\theta}, \hat{\pi}_{1+}, \hat{\pi}_{+1})$，这时，样本比例 $\hat{\pi}_{1+} = n_{1+}/n$ 以及 $\hat{\pi}_{+1} = n_{+1}/n$。关于 θ 的剖面似然置信区

间就是满足以下条件的一组 θ_0：

$$-2\left[L(\theta_0,\hat{\pi}_{1+}(\theta_0),\hat{\pi}_{+1}(\theta_0))-L(\hat{\theta},\hat{\pi}_{1+},\hat{\pi}_{+1})\right]<\chi_1^2(\alpha)。$$

这个区间包含了在名义水平为 α 的情况下所有未被似然比检验拒绝的 θ_0。

有些软件支持剖面似然法(如 SAS,参见附录 A 中的表 A.2)。在第 6.7.1 节我们将介绍一种与此有关的方法,即*条件似然函数*(*conditional likelihood function*),通过以充分统计量(sufficient statistics)为条件来消除冗余参数。当存在很多冗余参数时,这种方法非常有用。计分法和基于似然函数的区间的一个优点在于,与沃尔德区间不同,它们不会受到样本相对风险或发生比之比等于 0 或∞ 的影响。

在这节中我们已经讨论了区间估计。显著性检验通常对应的是对数发生比之比、对数相对风险或者比例之差等于零假设值 0.0。这些都是 2×2 表格满足独立性的特例。在下一节中,我们介绍关于二维列联表的独立性检验。

3.2 二维列联表的独立性检验

按照概率为 $\{\pi_{ij}\}$ 的多项分布进行抽样所形成的 $I\times J$ 列联表,其统计独立性检验的零假设为对于所有的 i 和 j 都满足 H_0: $\pi_{ij}=\pi_{i+}\pi_{+j}$。如果在 I 行中每行都是独立的多项分布样本,独立性意味着在各行之间每一种结果发生的概率都相同。这里讨论的是单一的多项分布样本,但同样的检验也适用于多个独立的多项分布样本。

3.2.1 皮尔逊和似然比卡方检验

在第 1.5.2 节中,我们介绍了多项概率检验的皮尔逊 X^2 统计量(式 1.15)。通过 X^2 进行的独立性检验使用 n_{ij}替代 n_i 以及 $\mu_{ij}=n\pi_{i+}\pi_{+j}$替代 μ_i。当 H_0 成立时,$\mu_{ij}=E(n_{ij})$。通常情况下,$\{\pi_{i+}\}$ 和 $\{\pi_{+j}\}$ 是未知的。它们的最大似然估计值等于样本的边际比例,即 $\hat{\pi}_{i+}=n_{i+}/n$ 以及 $\hat{\pi}_{+j}=n_{+j}/n$,因而,期望频数的估计值为 $\{\hat{\mu}_{ij}=n\hat{\pi}_{i+}\hat{\pi}_{+j}=n_{i+}n_{+j}/n\}$。这时,$X^2$ 等于

$$X^2=\sum_i\sum_j\frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}。\tag{3.10}$$

皮尔逊(Pearson,1900,1994,1922)认为,用估计值 $\{\hat{\mu}_{ij}\}$ 替代 $\{\mu_{ij}\}$ 不会影响 X^2 的分布。由于列联表包括 IJ 个类别,他指出 X^2 渐近服从自由度 df = $IJ-1$ 的卡方分布。事实上,由于 $\{\hat{\mu}_{ij}\}$ 需要对 $\{\pi_{i+}\}$ 和 $\{\pi_{+j}\}$ 进行估计,按照第 1.5.6 节,

$$\text{df}=(IJ-1)-(I-1)-(J-1)=(I-1)(J-1)。$$

$\{\pi_{i+}\}$ 和 $\{\pi_{+j}\}$ 的维度反映了限定条件 $\sum_i\pi_{i+}=\sum_j\pi_{+j}=1$。R. A. Fisher(1922)纠正了皮尔逊的这一错误(参见第 16.2 节),并首次引入了*自由度*(*degrees of freedom*)的概念(尽管皮尔逊已经分析了有关卡方分布族的指标,但是他并没有明确使用"自由度"这一概念)。

计分检验给出了 X^2 统计量,但用似然比检验会得出不同的值。对于多项分布抽样,似然函数的核函数为

$$\prod_i\prod_j\pi_{ij}^{n_{ij}},\quad 其中所有\ \pi_{ij}\geqslant 0\quad 且\quad \sum_i\sum_j\pi_{ij}=1。$$

当"H_0: 独立性"成立时,$\hat{\pi}_{ij}=\hat{\pi}_{i+}\hat{\pi}_{+j}=n_{i+}n_{+j}/n^2$。一般情况下,$\hat{\pi}_{ij}=n_{ij}/n$。因而,似然函

数之比等于

$$\Lambda = \frac{\prod_i \prod_j (n_{i+} n_{+j})^{n_{ij}}}{n^n \prod_i \prod_j n_{ij}^{n_{ij}}}。$$

似然比卡方统计量为 $-2 \log \Lambda$,由 G^2 来表示,也即

$$G^2 = -2 \log \Lambda = 2 \sum_i \sum_j n_{ij} \log(n_{ij}/\hat{\mu}_{ij}) \tag{3.11}$$

其中 $\{\hat{\mu}_{ij} = n_{i+} n_{+j}/n\}$。$G^2$ 和 X^2 的值越大,就有越多的证据拒绝独立性假设。

一般情况下,参数空间包含满足线性限定条件 $\sum_i \sum_j \pi_{ij} = 1$ 的 $\{\pi_{ij}\}$,其维度为 $IJ - 1$。在 H_0 下,$\{\pi_{ij}\}$ 由 $\{\pi_{i+}\}\{\pi_{+j}\}$ 决定,因此相应的维度为 $(I-1)+(J-1)$。二者的维度之差等于 $(I-1)(J-1)$。在大样本的情况下,G^2 服从自由度 $df = (I-1)(J-1)$ 的卡方零分布。因此,G^2 和 X^2 具有相同的极限卡方零分布。事实上,二者渐近等价,$X^2 - G^2$ 依概率收敛于零(参见第 14.3.4 节)。这一关于多项分布抽样的极限特性对其他的抽样方案同样成立(Roy and Mitra, 1956; Watson, 1959)。

这些结果适用于单元格数量固定的情况。这种情况下,随着 n 的上升,$\{\mu_{ij} = n\pi_{ij}\}$ 也会上升,$\{n_{ij}\}$ 的多项分布更近似于多元正态分布,进而 X^2 和 G^2 也就更接近卡方分布。与 G^2 相比,X^2 向卡方收敛的速度更快。当 $n/IJ < 5$ 时,G^2 的卡方近似通常较差。在 I 或 J 很大的情况下,即使有些期望频数小到 1 但大多数大于 5 时,对 X^2 的近似仍比较令人满意。这一问题,我们会在第 9.8.4 节进一步讨论。无论何种情况,当不确定 n 是不是足够大时,可以考虑使用小样本方法(第 3.5 节)。

3.2.2 例子:教育与宗教正统主义

表 3.2 给出了按照最高受教育水平划分的宗教信仰正统程度。该表也列出了在"H_0:独立性"下所估计的期望频数。例如,$\hat{\mu}_{11} = n_{1+} n_{+1}/n = (424 \times 886)/2\,726 = 137.8$。卡方统计量为 $X^2 = 69.2$,$G^2 = 69.8$,自由度 $df = (3-1)(3-1) = 4$。相应的 P 值 $< 0.000\,1$。这些统计量表明,最高受教育水平与宗教信仰正统程度之间存在非常显著的关联。

表 3.2 教育与宗教信仰

	宗教信仰			
最高受教育水平	正统派	中间派	开明派	小计
高中以下	178	138	108	424
	(137.8)[1]	(161.5)	(124.7)	
	(4.5)[2]	(−2.6)	(−1.9)	
高中或大专	570	648	442	1 660
	(539.5)	(632.1)	(488.4)	
	(2.6)	(1.3)	(−4.0)	
大学本科或研究生	138	252	252	642
	(208.7)	(244.5)	(188.9)	
	(−6.8)	(0.7)	(6.3)	
小计	886	1 038	802	2 726

来源:1996 年综合社会调查,美国民意研究中心(National Opinion Research Center)。1 独立性检验的期望频数;2 标准化皮尔逊残差。

3.3　对卡方检验的进一步分析

与所有显著性检验一样,独立性的卡方检验应用价值有限。P 值小表明存在很显著的关联,但是它没有给出任何关于关联的性质或强度的信息。统计学家很早就警告了仅仅依赖于卡方检验的结果而忽视分析关联本身的危险(如 Berkson, 1938;Cochran, 1954)。在这节中,我们讨论通过对卡方检验的进一步分析以获取更多关于关联的信息的方法。

3.3.1　皮尔逊残差和标准化残差

将每个单元格中的观测频数和估计的期望频数进行对比有助于了解关联的性质。在 H_0 下,较大的$(n_{ij}-\hat{\mu}_{ij})$差值一般出现在 μ_{ij}较大的单元格里。例如,对于泊松抽样,n_{ij} 也即$(n_{ij}-\mu_{ij})$的标准差等于 $\sqrt{\mu_{ij}}$;$(n_{ij}-\hat{\mu}_{ij})$的标准差小于$(n_{ij}-\mu_{ij})$的标准差,但它与 $\sqrt{\mu_{ij}}$成比例。因而,这种原始值的差异(raw difference)价值有限。单元格的*皮尔逊残差*(*Pearson residual*)被定义为

$$e_{ij}=\frac{n_{ij}-\hat{\mu}_{ij}}{\hat{\mu}_{ij}^{1/2}},\tag{3.12}$$

该残差对原始的差值进行了调整。皮尔逊残差与皮尔逊统计量的关系可表示为 $\sum_i\sum_j e_{ij}^2=X^2$。

在 H_0 假设下,$\{e_{ij}\}$渐近服从均值为 0 的正态分布。但是,它们的渐近方差小于 1.0,平均为$[(I-1)(J-1)]/$(单元格数),这一点我们将在第 14.3.2 节加以证明。将皮尔逊残差与标准正态分布中的百分位点进行比较会得出单元格拟合不足的保守估计。

将皮尔逊残差除以它的标准误可得到*标准化皮尔逊残差*(*standardized Pearson residual*),该残差渐近服从标准正态分布(Haberman, 1973a;另见第 14.3.2 节)。对于"H_0: 独立性"假设,标准化皮尔逊残差等于

$$\frac{n_{ij}-\hat{\mu}_{ij}}{[\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})]^{1/2}}。\tag{3.13}$$

如果标准化皮尔逊残差的绝对值超过 2 或 3,表明在相应单元格 H_0 的拟合不充分。自由度越大,越有可能出现较大的标准化皮尔逊残差,从纯随机的角度来说,出现至少一个很大的残差值的可能性也更大。

3.3.2　例子:对教育和宗教正统主义的再讨论

表 3.2 还给出了独立性检验的标准化皮尔逊残差。例如,$n_{11}=178$,$\hat{\mu}_{11}=137.8$。相应的边际比例为 $p_{1+}=424/2\,726=0.156$ 和 $p_{+1}=886/2\,726=0.325$。按照公式 3.13,这个单元格的标准化皮尔逊残差等于

$$(178-137.8)/[(137.8)(1-0.156)(1-0.325)]^{1/2}=4.5。$$

这个单元格所显示的 n_{11}和 $\hat{\mu}_{11}$之间的差异,比在两个变量真正独立的假设下所期望的差异大得多。

表 3.2 表明,对于受教育水平低于高中且信仰正统派宗教的研究对象,以及受教育水平为本科或研究生且信仰开明派的研究对象,都存在很大的正残差。这意味着,落在

这些单元格中的观测值明显比“H_0: 独立性”所预测的更多。相似地,具有高受教育水平与正统派信仰,以及低受教育水平与开明派信仰的对象,其观测值则比独立性假设所预测的少。

这一趋势可以通过发生比之比来描述。由表 3.2 的第一行和最后一行以及第一列和最后一列所组成的 2×2 表格的样本发生比之比为$(178\times252)/(108\times138)=3.0$。对于拥有本科或研究生学位的对象而言,选择开明派而不是正统派信仰的发生比是那些教育水平在高中以下的对象的 3.0 倍。

3.3.3 卡方的分割

令 Z 表示一个服从标准正态分布的随机变量,则有 Z^2 服从自由度 $df=1$ 的卡方分布。自由度 $df=v$ 的卡方随机变量可以表示为 $Z_1^2+\cdots+Z_v^2$,其中 $Z_1,\cdots,Z_v$ 为相互独立的标准正态变量。同样,自由度 $df=v$ 的卡方统计量可以分割成独立的卡方项——比如 v 个自由度 $df=1$ 的项。相反,如果 X_1^2 和 X_2^2 是自由度分别为 v_1 和 v_2 的独立的卡方随机变量,那么 $X^2=X_1^2+X_2^2$ 服从自由度 $df=v_1+v_2$ 的卡方分布。另外,可以将卡方检验统计量进行分割,以使不同的项代表不同的效应。这种分割或许可以揭示,某个关联主要是因为哪些类别或组别之间的差异引起的。

我们首先介绍对 $2\times J$ 表格独立性检验的分割。将自由度 $df=(J-1)$ 的 G^2 分割为 $J-1$项,其中第 j 项对应于将前 j 列合并成第 1 列和第 $j+1$ 列组成的一个 2×2 表格的 G^2。也就是说,$2\times J$ 表格独立性检验的 G^2,等于一个比较前两列的统计量,加上合并前两列并将其与第三列比较的统计量,以此类推,一直到合并前 $J-1$ 列并将其与最后一列比较的统计量(关于这个分割的证明,将在第 9.2.4 节给出)。其中,每一项统计量的自由度均为 $df=1$。

更为直观的方法是分别计算$(J-1)$个由每列与某一特定列,比如最后一列,所组成的 2×2 表格的 G^2。然而,由此形成的统计量之间并不是相互独立的,它们的加总也不等于整个表格的 G^2(这在目前超出了我们讨论的范围,但是它与形成两个非正交表格的对数发生比之比的对数概率有关)。

对于 $I\times J$ 表格,独立的卡方项来自于比较第一列和第二列,然后将其合并与第三列进行比较,……,以此类推。这 $J-1$ 个统计量的自由度均为 $df=I-1$。更进一步的分割包含$(I-1)(J-1)$个自由度为 $df=1$ 的统计量。这样的分割(Lancaster, 1949)适用于$(I-1)(J-1)$个单独的 2×2 表格:

$$\begin{array}{c|c} \sum\limits_{a<i}\sum\limits_{b<j} n_{ab} & \sum\limits_{a<i} n_{aj} \\ \hline \sum\limits_{b<j} n_{ib} & n_{ij} \end{array}, \qquad (3.14)$$

其中 $i=2,\cdots,I$ 以及 $j=2,\cdots,J$。有关其他的分割方法,参见 Gilula 和 Haberman(1998)以及 Goodman(1969a,1971b)。

3.3.4 例子:精神分裂症的病因

表 3.3 给出的是关于精神病学家的一个样本,按照他们的理论学派及其对精神分裂症病因的观点进行了划分。这里,$G^2=23.04$,自由度为 $df=4$。为了更好地理解这一关联,我们将 G^2 分割成四个独立的项。表 3.4 给出了按照式 3.14 进行分割所得到的子表。

表 3.3 主流精神病理论学派对精神分裂症病因的看法

精神病理论学派	精神分裂症病因		
	生物遗传	环境	综合
折中派	90	12	78
医学派	13	1	6
精神分析学派	19	13	50

来源:授权重印,数据基于:B. J. Gallagher Ⅲ, B. J. Jones, and L. P. Barakat, *J. Clin. Psychol.* 43:438-443 (1987)。

表 3.4 对表 3.3 进行卡方分割的子表

	Bio	Env		Bio + Env	Com		Bio	Env		Bio + Env	Com
Ecl	90	12	Ecl	102	78	Ecl + Med	103	13	Ecl + Med	116	84
Med	13	1	Med	14	6	Psy	19	13	psy	32	50

第一个子表比较折中派和医学派的精神病学家关于精神分裂症为生物遗传还是环境因素所导致的观点,给定其观点为两种之一。对于这个子表,$G^2=0.29$,自由度为 df = 1。第二个子表比较这两个学派的精神病学家将病因归结于综合因素而不是生物遗传或环境的比例。这个子表的 $G^2=1.36$,自由度为 df = 1。将这二项求和等于对表 3.3 中的前两行进行独立性检验的 G^2。没有证据表明折中派和医学派的精神病学家对精神分裂症的病因持有不同的看法。

接下来我们将折中派与医学派合并,然后将它们与精神分析学派进行比较。表 3.4 的第三个子表比较它们关于(生物遗传,环境因素)的划分,给出的 $G^2=12.95$,自由度为 df = 1。第四个子表比较它们关于(生物遗传或环境,综合因素)的划分,$G^2=8.43$,自由度为 df = 1。

与其他学派相比,精神分析学派似乎更倾向于将精神分裂症归结为综合因素。在那些认为是生物遗传或者环境因素的精神病学者中,精神分析学派在一定程度上比其他学派更倾向于选择环境因素。这四个 G^2 项之和等于对整个表格进行独立性检验的 G^2 值 23.04。

3.3.5 分割的原则

Goodman(1968,1969a,1971b)和 Lancaster(1949,1969)讨论了决定独立卡方项的原则。对于构建子表而言,其必要条件为:

1. 所有子表的自由度相加必须等于总表的自由度。
2. 每个总表中的单元格计数必须且只能是一个子表中的单元格计数。
3. 每个总表的边际总计必须且只能是一个子表中的边际总计。

如果在某种分割方案中,其自由度的加总满足条件,但是 G^2 的总和不相等,那么分割的各项之间不独立。

G^2 统计量可以进行精确的分割。皮尔逊的 X^2 则不一定等于所有子表的 X^2 加总。可以对这些子表分别使用 X^2 统计量,只是它们不必然等于对总表 X^2 的精确的代数分割。不过,当零假设都成立时,X^2 确实与 G^2 渐近等价。另外,当表格中某些计数很小时,在大样本卡方检验中使用 X^2 来分析子表会更安全。

3.3.6 卡方检验的局限性

独立性的卡方检验仅仅反映是否存在关联。这对于回答数据分析的所有问题来说是远远不够的。除检验结果外，我们需要对关联本身进一步加以研究：分析残差，将卡方进行分割，以及估计诸如发生比之比等描述关联强度的参数。

卡方检验在适用的数据类型方面也具有局限性。例如，它要求规模较大的样本。同时，由于 X^2 和 G^2 中所使用的 $\{\hat{\mu}_{ij}=n_{i+}n_{+j}/n\}$ 取决于表格的边际总计但不取决于行或列之间的排序，因此，X^2 和 G^2 不会因为对行或列进行随意的重新排列而发生改变。这意味着，它们将划分尺度视为名义尺度。当至少有一个变量是定序变量时，利用排序信息的检验统计量往往更合适。我们将在第 3.4 节介绍这样的检验。

3.3.7 为什么要考虑独立性?

现实情况下，诸如独立性这样的理想结构往往并不成立。在大样本情况下，如表 3.2，很容易得到一个很小的 P 值。那么，我们为什么要花精力去考虑独立性作为联合分布的一种可能结果呢？原因之一是出于对简约模型的偏好。如果独立性模型能对真实的概率进行很好的近似，那么除非 n 非常大，独立性模型对单元格概率的估计 $\{\hat{\pi}_{ij}=n_{i+}n_{+j}/n^2\}$ 往往优于样本比例 $\{p_{ij}=n_{ij}/n\}$。独立性假设对应的最大似然估计修匀了样本计数，在一定程度上消除了随机抽样波动的影响。

均方差(mean-squared error, MSE)的公式

$$\text{MSE}=方差+(偏差)^2$$

解释了为什么独立性模型估计值会具有较小的均方差。尽管有可能存在偏差，由于它们是基于对较少的参数($\{\pi_{i+}\}$ 和 $\{\pi_{+j}\}$ 而不是 $\{\pi_{ij}\}$)的估计，因而方差较小。因此，除非 n 非常大，以至于偏差项比方差大得多，独立性估计值均方差会比较小。

我们通过表 3.5 来对此加以详细说明，其中对于 $\pi_{i+}=\pi_{+j}=\dfrac{1}{3}$，有 $\pi_{ij}=\pi_{i+}\pi_{+j}[1+\delta(i-2)(j-2)]$，这里 $-1<\delta<1$，$\delta=0$ 时意味着独立性。当 δ 接近于零时，独立性模型能够对这一关系进行很好的近似。两个估计值的总和均方差分别为：

$$\begin{aligned}\text{MSE}(\{p_{ij}\}) &= \sum_i\sum_j E(p_{ij}-\pi_{ij})^2 = \sum_i\sum_j \text{var}(p_{ij})\\ &= \sum_i\sum_j \pi_{ij}(1-\pi_{ij})/n = (1-\sum_i\sum_j \pi_{ij}^2)/n,\\ \text{MSE}(\{\hat{\pi}_{ij}\}) &= \sum_i\sum_j E(\hat{\pi}_{ij}-\pi_{ij})^2。\end{aligned}$$

表 3.5 进行估计值比较的单元格概率

$(1+\delta)/9$	1/9	$(1-\delta)/9$
1/9	1/9	1/9
$(1-\delta)/9$	1/9	$(1+\delta)/9$

对于表 3.5，

$$\text{MSE}(\{p_{ij}\})=\frac{1}{n}\left\{\frac{8}{9}-\frac{4\delta^2}{81}\right\},$$

通过相当繁琐的计算得出，

$$\mathrm{MSE}(\{\hat{\pi}_{ij}\}) = \frac{1}{n}\left\{\frac{4}{9}+\frac{4}{9n}\right\}+\frac{4\delta^2}{81}\left\{1-\frac{2}{n}+\frac{2}{n^2}-\frac{2}{n^3}\right\}。$$

表3.6列出了在 δ 和 n 的不同取值下总和均方差的值。当 $\delta=0$ 时，$\mathrm{MSE}(\{p_{ij}\})=8/9n$；然而，对于大样本 n，$\mathrm{MSE}(\{\hat{\pi}_{ij}\})\approx 4/9n$。这时，独立性估计值明显好于样本比例。当表格接近于独立（$\delta\approx 0$）并且 n 不是很大时，独立性估计值的均方差仅仅约为样本比例的一半。当 $\delta\neq 0$ 时，$\{\hat{\pi}_{ij}\}$ 不具有一致性，当 $n\to\infty$ 时，$\mathrm{MSE}(\{\hat{\pi}_{ij}\})\to 4\delta^2/81$（而 $\mathrm{MSE}(\{p_{ij}\})\to 0$）。但是，当表格接近于独立时，即使对于较大的 n（如当 $n=500$ 以及 $\delta=0.1$ 时），独立性估计值的总和均方差也较小。

表3.6　比较样本比例和独立性估计值的总和均方差（×10 000）

	$\delta=0$		$\delta=0.1$		$\delta=0.2$		$\delta=0.6$		$\delta=1.0$	
n	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$
10	889	489	888	493	887	505	871	634	840	893
50	178	91	178	95	177	110	174	261	168	565
100	89	45	89	50	89	65	87	220	84	529
500	18	9	18	14	18	28	17	186	17	500
∞	0	0	0	5	0	20	0	178	0	494

3.4　定序变量的二维表格

在对具有定序尺度的表格进行独立性检验时，X^2 和 G^2 的卡方检验忽略了排序的信息。当行之间和/或列之间存在排序时，通常存在更有效的检验方法。

3.4.1　与独立性相对应的线性趋势

当行变量 X 和列变量 Y 是定序变量时，关联往往具有正向或负向的趋势。本节稍后将讲到的一种统计推断方法，就是利用这种定序测度的单调趋势。更常见的一种分析方法是，将各个类别进行赋值，然后描述这种线性趋势（*linear trend*）。

对正向或负向线性趋势灵敏的检验统计量与相关系数有关。令 $u_1\leqslant u_2\leqslant\cdots\leqslant u_I$ 为不同行的赋值，令 $v_1\leqslant v_2\leqslant\cdots\leqslant v_J$ 为列的赋值。这些值具有与类别相同的排序，它们给出了各个类别之间的距离，实际上将测量尺度视为定距尺度，即赋值相差较大的类别之间相距较远。

总和 $\sum_i\sum_j u_i v_j n_{ij}$ 是将行和列的赋值相乘，并按照相应的频数加权。该统计量的值与 X 和 Y 的协变关系（covariation）有关。对于给定的赋值，X 和 Y 的相关系数 r 等于将上面的总和统计量在 -1 到 $+1$ 的水平上进行标准化（实际上，当对 n 个研究对象的 X 和 Y 两组赋值都进行均值为0、标准差为1的线性转换后，r 便等于上面的总和）。相关系数的绝对值越大，两个变量在线性维度上偏离独立性就越远。

对于备择假设为真实相关系数不为零，双边独立性检验的统计量是

$$M^2=(n-1)r^2。\tag{3.15}$$

这个统计量会随着 $|r|$ 或 n 的上升而上升。在大样本情况下，它近似服从自由度 $\mathrm{df}=1$ 的卡方分布（Mantel, 1963）。与 X^2 和 G^2 相同，该统计量取很大的值意味着独立性不成立，

P 值为大于所观测到的值的右尾概率。P 值小并不意味着这个关联就是线性的，它仅仅表明这个关联中的一个线性成分有助于构建拒绝 H_0 的统计效能。在这个检验中，两个变量的关系是对称的。

3.4.2 例子：对工作满意度的再讨论

表 2.8 给出了 96 个研究对象的工作满意度和收入的分布。常见的独立性检验卡方统计量是 $X^2=6.0$ 和 $G^2=6.8$，自由度为 df = 9（P 值分别为 0.74 和 0.66）。这些统计量表明，表中的两个变量之间不存在关联，但这些检验忽略了行和列的排序。对工作满意度赋值为(1,2,3,4)，并取各组收入的中值(7.5,20,32.5,60)，这时两个变量的相关系数为 $r=0.200$。线性趋势检验统计量 $M^2=(96-1)(0.200)^2=3.81$，表示两个变量之间存在一定程度的关联（$P=0.051$）。相应的单边（正向趋势）检验的结果更为显著，统计量为 $M=\sqrt{n-1}\,r=1.95$（$P=0.026$）。

由于 X^2 和 G^2 的值较小，因而得出显著的正关联的结果似乎有些意外。当存在正向或负向关联趋势时，在分析过程中考虑这种趋势比忽略相应信息所得出的 P 值小得多。

3.4.3 与独立性相对应的单调趋势

定序变量并不存在一个特定的度量标准。检测与独立性相矛盾的线性趋势需要对 X 和 Y 进行赋值，并将其视为定距变量。除此之外，还可以进行一种严格的定序分析，通过定序关联的度量指标如 γ(gamma)来分析较弱的单调性假设（第 2.4.4 节）。

在随机大样本的情况下，样本 γ 统计量近似服从正态分布。其标准误可以通过 δ 方法计算求得（习题 3.27）。γ 是通过检验统计量 $z=\hat{\gamma}/\text{SE}$ 对定序变量进行独立性检验的基础。它的置信区间可以用来描述正向或负向单调关联的强度。

对于表 2.8 中收入和工作满意度的数据，在第 2.4.5 节我们求得了 $\hat{\gamma}=0.221$。这表明，两个变量之间存在较弱的关联趋势，在较高的收入水平上工作满意度也较高。统计软件（如 SAS 的 PROC FREQ）给出的 γ 的标准误为 0.117。这样，$z=0.221/0.117=1.89$（单边检验中 $P=0.03$），有一定的证据支持 $\gamma>0$。γ 的 95% 的置信区间约为 $0.221\pm1.96(0.117)$，即(−0.01,0.45)。因而，收入和工作满意度之间最多存在中等强度的正向关联。

3.4.4 定序检验的额外效能

利用 X^2 和 G^2 所进行的检验是独立性检验的最一般情形，这些检验回答单元格概率之间是否存在任何形式的统计相依的问题。检验的自由度$(I-1)(J-1)$表明，备择假设比零假设多$(I-1)(J-1)$个参数——用来描述关联的非冗余的发生比之比（参见式 2.10）。这些统计量可用于检测参数之间所存在的任意模式。为了保证这种一般性，这些检验牺牲了检测某种具体模式的灵敏性。

与之相反，定序变量的列联表分析试图通过一个参数来描述关联关系。例如，M^2 使用了相关系数。当卡方检验统计量只有一个参数时（如 M^2 或$(\hat{\gamma}/\text{SE})^2$），它的自由度为 df = 1。这样，当两个变量之间确实存在正向或负向关联趋势时，定序检验比使用 X^2 和 G^2 的检验具有更强的统计效能。由于自由度等于卡方分布的均值，自由度为 df = 1 的较大的 M^2 会比一个大小相当而自由度为 df = $(I-1)(J-1)$的 X^2 和 G^2 落到相应分布的右

尾更远的地方,也即其所对应的 P 值更小。定序检验与普通检验在效能上的潜在差别随着 I 和 J 的增加而增大。在第 6.4 节,我们将介绍这种效能比较背后所隐含的理论。

3.4.5 赋值的选取

通常情况下,我们并不清楚如何对需要赋值的统计量,如 M^2,进行赋值。Cochran (1954)写道:"任何事先没有考虑对结果的影响的赋值都会给出有效的检验。但如果某组赋值很糟糕、严重扭曲了定序划分实际所隐含的数值刻度,相应的检验就不灵敏。因此,赋值应当体现出对相应划分的构建和使用方式的最佳认识。"理想情况下,可以由专家达成的共识来决定赋值的尺度,并在之后的解释中使用同一尺度。

那么,分析结果在多大程度上会受到不同赋值选择的影响呢?这个问题并没有简单的答案,但是不同的赋值体系能够导致非常不同的结果(如 Graubard and Korn, 1987)。对于大多数的数据来说,选取不同的单调赋值会给出相似的结果。两组互为线性转换的赋值,比如(1,2,3,4)和(0,2,4,6),具有完全相同的相关系数,因而它们的 M^2 也相等。然而,如果数据分布高度不均,即某些类别的观测值比其他类别多很多,那么结果可能会因赋值的不同而不同。

表 3.7 展示了这种可能性。表中所示为一项有关母亲饮酒与先天畸形的前瞻性研究。在怀孕三个月后,样本中的妇女完成了一份关于饮酒状况的问卷。在孩子出生后,对是否存在先天性性器官畸形进行了记录。当一个变量是定类变量但只有两类时,在统计上可将其视为定序变量来使用。例如,我们可以将畸形视为定序变量,以"出现畸形"为"高","未出现"为"低"。因为只存在两行,任何一组关于行的不等赋值都是另一组不同赋值的线性转换,相应的 M^2 的值相同。饮酒行为,由每天的平均饮酒量来测量,是一个定序的解释变量。它的取值本质上是连续的,我们首先使用赋值$\{v_1=0, v_2=0.5, v_3=1.5, v_4=4.0, v_5=7.0\}$,其中最后一个赋值具有一定的随意性。这组赋值所对应的 $M^2=6.57$,P 值为 0.010。与之相对,如果使用间距相等的赋值(1,2,3,4,5),$M^2=1.83$,由此得出的结论要弱得多($P=0.18$)。

表 3.7 检验结果受赋值影响的例子

	饮酒量(每天平均饮酒数)				
畸形	0	<1	1~2	3~5	≥6
不存在	17 066	14 464	788	126	37
存在	48	38	5	1	1

来源:由生物计量学学会(Biometric Society, Graubard and Korn 1987)授权重印。

另一种赋值的方法是利用数据中类别的排序自动生成赋值。将所有样本从 1 到 n 进行排序,然后使每一类别的赋值等于它所包括的对象的平均排序,也称作*中位秩*(*midranks*)。例如,表 3.7 所对应的数据中,17 114 个饮酒水平为 0 的对象所对应的排序为 1 到 17 114。这些排序的中位秩是$(1+17\ 114)/2=8\ 557.5$。同样地,后四个类别的中位秩分别为 24 365.5, 32 013, 32 473 和32 555.5。基于中位秩赋值的 $M^2=0.35$,所得结论仍然较弱($P=0.55$)。

为什么会出现这种情况呢?观测值较少的相邻类别必然具有相近的中位秩。上例中,由于后三个类别与前两个类别相比观测值很少,它们的中位秩更为接近。这种赋值方法将饮酒水平为1-2 杯(第3 类)视为与≥6 杯(第 5 类)比与0 杯(第1 类)相比更为接近。这样赋值并不恰当。通常,选取那些能够反映类别间的距离的赋值会好一些。当我

们对距离不确定时，应当进行灵敏度分析（sensitivity analysis），即选取两到三种不同的赋值来检查结论是否一致。当类别本身无法给出明显的赋值选择时，如政治态度的分类（激进、温和、保守），使用间距相等的赋值往往是一个合理的折中选择。

如果 X 和 Y 都是定序变量且 M^2 使用的是中位秩赋值，那么计算 M^2 时所使用的相关系数被称为*斯皮尔曼的*ρ（*Spearman's rho*）。

3.4.6 关于 $I\times 2$ 和 $2\times J$ 表格的趋势检验

当 I 或 J 等于 2 时，基于线性或单调趋势的检验就简化为广为应用的一般情况。X 为二分变量的 $2\times J$ 表格通常用于对两组进行比较的情况，如行变量代表两种干预方式。用 $\{u_1=0, u_2=1\}$ 来对 X 的类别进行赋值，对应的 M^2 中的协变指标 $\sum_i\sum_j u_i v_j n_{ij}$ 就简化为 $\sum_j v_j n_{2j}$。它对第 2 行中所有对象的 Y 值进行加总。该值除以第 2 行中所有对象的总数，便求得这一行的平均值。事实上，这里就是通过 M^2 发现两行中 Y 的平均取值的差异。

当 Y 使用中位秩赋值时，对 $2\times J$ 表格的 M^2 检验会对两行的平均排序的差异相当敏感。这个检验被称为 *Wilcoxson 或 Mann-Whitney 检验*（*Wilcoxon or Mann-Whitney test*）。大多数统计学教科书在介绍完全排序的结果变量的非参数检验时会提及这一检验。而 $2\times J$表格对应的是对该检验的扩展，即在 Y 存在相同取值时使用其中位秩。在大样本情形下，该非参数检验使用服从标准正态分布的 z 统计量。该统计量的平方与使用任意行赋值和中位秩列赋值时的 M^2 等价。同时，它也与基于相协对和相异对数量的检验统计量（如 γ）渐近等价。

当 Y 具有两个类别时，对应的表格为 $I\times 2$。这时，线性趋势统计量是用以检验结果变量的每个类别发生概率的线性趋势，比如出现畸形的概率作为饮酒量的函数。在这种情况下，该检验常常被称为 *Cochran-Armitage 趋势检验*（*Cochran-Armitage trend test*），这一检验我们将留在第 5.3.5 节加以介绍。

3.4.7 定类—定序表格

当两个变量都是定序变量时，我们可以使用相关系数或 γ 进行检验。当其中一个变量是定类变量且其类别多于两类时，我们需要用到其他的统计量。这些统计量往往是通过对在定类变量的不同类别对应的定序变量的均值变动来进行检验。对这种情形的讨论，可参见第 7.5.3 节、注解 3.6 以及习题 3.28。

3.5 小样本的独立性检验

本章前四节所介绍的统计推断方法都是针对大样本的情况。在小样本的情况下，可以使用精确小样本分布，而不是通过大样本的近似来进行统计推断。在这节中，我们介绍小样本的独立性检验，其中首先将介绍 R. A. Fisher 提出的关于 2×2 表格的检验。

3.5.1 关于 2×2 表格的 Fisher 精确检验

在第 3.5.7 节，我们所介绍的分布不依赖于未知参数，但受制于相应列联表的边际总计。这些边际总计通常并不是事先给定的。例如，在泊松分布抽样中，没有什么是给定的；在多项分布抽样的情况下，只有 n 是给定的；而对于行间相互独立的二项分布抽样，只有两个行的边际总计是给定的。在上述情况中，如果行和列的边际总计都给定，那

么“H_0: 独立性”对应着超几何分布(hypergeometric distribution)。

$$P(t) = P(n_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1} - t}}{\binom{n}{n_{+t}}}。\quad (3.16)$$

该公式表示$\{n_{ij}\}$的分布只取决于n_{11}。给定边际总计,n_{11}决定了其他三个单元格计数。n_{11}的可能取值范围为$m_- \leq n_{11} \leq m_+$,在这里$m_- = \max(0, n_{1+} + n_{+1} - n)$,$m_+ = \min(n_{1+}, n_{+1})$。

对于2×2表格,独立性等价于发生比之比$\theta = 1$。检验H_0: $\theta = 1$,P值就可以表示为多个超几何分布概率的求和。举例来说,考虑H_a: $\theta > 1$。对于给定的边际总计,n_{11}的值较大时对应的样本发生比之比也较大,检验结果也就更有可能支持H_a。这时,P值等于$P(n_{11} \geq t_o)$,其中t_o为n_{11}的观察值。这个关于2×2表格的检验被称为*Fisher 精确检验*(*Fisher's exact test*)(Fisher, 1934, 1935a,c;Irwin, 1935;Yates, 1934)。

3.5.2 Fisher 的品茶实验

R. A. Fisher(1935a)描述了他在一个位于北伦敦的农业研究实验室——洛桑(Rothamsted)实验基地——工作时的以下实验。Fisher 的同事,Muriel Bristol 宣称她在喝茶时能分辨出在杯中是先加的牛奶还是先加的茶叶(她喜欢先放牛奶)。为了对此进行测试,Fisher 让她品尝了八杯茶,其中的四杯是先加牛奶,另外四杯是先加茶。她被告知两种泡茶方式各有四杯,然后将茶杯随机递给她,要求她指出哪四杯是先加的牛奶。

表3.8给出了一种可能的结果。假如对泡茶方式的分辨结果要好于纯粹的猜测,则有$\theta > 1$,反映在泡茶方式与预测结果之间存在正向关联。对此,我们进行零假设为H_0: $\theta = 1$,备择假设为H_a: $\theta > 1$的 Fisher 精确检验。

表3.8 Fisher 的品茶实验

	猜测结果		
实际泡茶方式	先放牛奶	先放茶叶	小计
先放牛奶	3	1	4
先放茶叶	1	3	4
小计	4	4	

来源:基于 Fisher(1935a)所描述的实验。

实验设计本身给定了两个边际分布,Bristol 博士需要预测哪四杯茶先添加了牛奶。因此,n_{11}的零分布自然服从超几何分布。Fisher 精确检验的P值等于表3.8以及具有更多支持证据的表格发生的概率之和。在观察到的表格中,对先添加牛奶的茶杯的正确选择为$t_o = 3$次,其零分布概率等于

$$\frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = 0.229。$$

唯一一个更支持H_a的表格是$n_{11} = 4$,它的概率为0.014。所以,$P(n_{11} \geq 3)$所对应的P值为0.243。该结果并没有明确支持实际泡茶方式与预测之间存在关联。对于如此小的样本而言,这是很困难的。据 Fisher 的女儿(Box, 1978, p. 134)所言,在现实中 Bristol 确实令 Fisher 信服了她的这一能力。

3.5.3 Fisher 精确检验的双边 P 值

对于单边检验来说，将表格按照较大的 n_{11}、较大的发生比之比或较大的比例之差进行排列都可以得出相同的 P 值(Davis, 1986a)。而对于双边检验，使用不同的标准会给出不同的 P 值。

在计算双边检验的 P 值时，一种常用的方法是对所有满足 $p(t) \leqslant p(t_o)$ 的计数 t，加总公式3.16中的 $P(n_{11}=t)$，即对于观察到的值 t_0，相应 P 值为 $P=P[p(n_{11}) \leqslant p(t_o)]$。另一种方法是对偏离 H_0 的表格对应的 $p(t)$ 求和，也即，

$$P = P[\,|n_{11} - E(n_{11})| \geqslant |t_o - E(n_{11})|\,],$$

其中超几何分布的 $E(n_{11}) = n_{1+}n_{+1}/n$。这与观察到的皮尔逊统计量 X_0^2 对应的 $P(X^2 \geqslant X_0^2)$ 完全一样。第三种方法使用 $P=2\min[P(n_{11} \geqslant t_o), P(n_{11} \leqslant t_o)]$，但这种方法计算的 P 值有可能大于1。第四种方法使用 $P=\min[P(n_{11} \geqslant t_o), P(n_{11} \leqslant t_o)]$ 加上一个在另一尾部接近于但不大于该单尾概率的可得概率。

以上方法各有优劣(Blaker, 2000; Davis, 1986a; Dupont, 1986; Lloyd, 1988b; Mantel, 1987b; Yates and discussants, 1984)。由于分布的离散性和可能的偏斜问题，不同方法可能会给出不同的结果。将表格按照某种偏离 H_0 的距离排序的方法，如 X^2，可以自然地扩展到 $I \times J$ 表格的情况。

在应用中，双边检验要比单边检验更为常见。部分原因是因为运用双边检验的研究者可以避免被指责试图给出与其所期望的方向相一致的效应。当我们要进行一个在0.05水平上的检验时，如果我们可以认定这个效应的方向是确定的，那么最安全的做法是在0.025 水平上进行单边检验，从而可以避免相应的批评。例如，在 1998 年的文献《*临床试验的生物统计学原理*(*Biostatistical Principles for Clinical Trials*)》中，国际协调会议(the International Conference on Harmonization, ICH E9)指出："在可控的条件下，将单边检验中第 I 类错误的水平设定为相应双边检验的第 I 类错误的一半更为可取。这样，在估计两种干预方式效果的差异时，这种做法得出与常用的双边置信区间相一致的结果。"

3.5.4 离散性与检验结果的保守性问题

在小样本情况下，公式 3.16 的超几何分布是高度离散的，这时 n_{11} 以及与之相对应的 P 值只能取相对很少的有限个值。这种情况下，往往不太可能正好得出某一给定的显著性水平，如 0.05。

例如，在品茶实验中，n_{11} 仅可以取 4, 3, 2, 1, 0。与之相对应，单边检验的 P 值只能取 0.014, 0.243, 0.757, 0.986 和 1.0。如果我们在 P 值不超过 0.05 时拒绝 H_0，那么 0.05 并不是发生第 I 类错误的概率。上述 P 的可能取值中，只有 0.014 没有超过 0.05，这样，在 H_0 为真的情况下错误地拒绝它的概率是 0.014，而不是 0.05。从这个意义上讲，传统的假设检验方法所得结论太保守：第 I 类错误的真实概率小于其名义水平。

通过在临界区域的边界进行与数据无关的随机化，可以求得任意给定的置信水平，从而决定是否拒绝 H_0。以品茶实验为例，假定当 $n_{11}=4$ 时我们拒绝 H_0，当 $n_{11}=3$ 时我们以 0.157 的概率拒绝 H_0，其他情况下我们接受 H_0；也即，当 $n_{11}=3$ 时，我们生成一个均匀随机变量 U，其取值为[0,1]，这样当 $U<0.157$ 时就拒绝 H_0。就 n_{11} 的超几何零分布期望值而言，置信水平等于

$$P(\text{拒绝 } H_0) = E\left[P(\text{拒绝 } H_0 \mid n_{11})\right]$$

$$= 1.0(0.014) + 0.157(0.229) + 0.0 \times P(n_{11} \leqslant 2) = 0.05。$$

通过随机化的扩展，Tocher(1950)表明 Fisher 检验是一致最大效能无偏检验(uniformly most powerful unbiased, UMPU)。

在应用中，进行与数据本身完全无关的随机化往往是不能接受的。我们建议仅报告 P 值。为了降低保守性问题的影响，可以报告中位 P 值(第 1.4.5 节)。该检验不再保证真实的 P(第Ⅰ类错误)不超过名义水平，但是在应用中它极少会明显超过名义水平。就品茶数据的单边检验而言，

$$\text{中位 } P \text{ 值} = (1/2)P(n_{11} = 3) + P(n_{11} > 3) = 0.129。$$

3.5.5　小样本的无条件独立性检验*

在比较二分结果变量在两组之间的差异时，一种常见的抽样假定是每一行都是独立的二项分布样本。这时，只有$\{n_{i+}\}$是给定的。对于泊松分布和多项分布的抽样设计，所有的边际分布都不是给定的。在这些情况下，以*两组(行和列)*边际计数为条件进行检验并不恰当。这里我们介绍另一种小样本检验，它根据独立的二项分布样本的特点，仅以行总计为条件进行检验。

令列联表中第 i 行来自参数为 π_i 的二项分布抽样，使用某检验统计量 T(如皮尔逊 X^2)来检验 H_0: $\pi_1 = \pi_2$。在$\{n_{i+}\}$给定的情况下，T 可以取一组离散的值，其中之一便是观察到的值 t_o。给定 $\pi_1 = \pi_2 = \pi$，P 值为 $P_\pi(T \geqslant t_o)$，可以由两个二项分布概率密度函数的乘积求得，也即满足 $T \geqslant t_o$ 的二项分布样本的组合中相应二项分布概率的乘积之和。由于 π 是未知的，实际的 P 值被定义为

$$P = \sup_{0 \leqslant \pi \leqslant 1} P_\pi(T \geqslant t_o)。$$

这就是*无条件的*小样本独立性检验。与 Fisher 精确检验一样，该检验的真实置信水平不会高于名义水平(例如，如果我们在 $P \leqslant 0.05$ 时拒绝 H_0，那么实际的 P(第Ⅰ类错误)不会大于 0.05)。

下面，我们以 2×2 表格的检验统计量 X^2 为例来加以说明。假设表格中分行的频数为(3,0/0,3)，其中行总计{3,3}是给定的，这也是二项分布抽样的样本规模。这时，样本的 $X^2 = 6.0$。该观测表格和表格(0,3/3,0)的 X^2 值是 X^2 的最大可能取值。对于任意满足 $\pi_1 = \pi_2$ 的 π 值，上述第一个表格出现的概率是$[\pi^3(1-\pi)^0][\pi^0(1-\pi)^3] = \pi^3(1-\pi)^3$(在第一行 3 次成功，0 次失败；在第二行 0 次成功，3 次失败)，即两个二项分布概率的乘积。类似地，第二个表格出现的概率为$(1-\pi)^3\pi^3$。因此，P 值等于 $P_\pi(X^2 \geqslant 6) = 2\pi^3(1-\pi)^3$，即相应两个表格的二项分布概率乘积之和。在 $0 \leqslant \pi \leqslant 1$ 的条件下，该 P 值的上确界(supremum)发生在 $\pi = \frac{1}{2}$时，这时 P 值为 $2(0.5)^3(0.5)^3 = 0.031$。相比之下，Fisher 精确检验的双边 P 值等于 $2\binom{3}{0}\binom{3}{3}\Big/\binom{6}{3} = 0.100$。

比较二项分布参数的无条件检验最早由 Barnard(Barnard,1945,1947)提出，尽管后来他(1949)为了支持 Fisher 精确检验对此进行了反驳。此后，其他几位研究者又提出了相应的检验方法(如 Haber, 1986;Suissa and Shuster, 1985)。

3.5.6 条件检验与无条件检验的对比*

在 Barnard 引入了无条件检验后，统计学家就如何正确地进行 2×2 表格的小样本分析展开了争论。Fisher 对无条件检验方法提出了批评，认为与观察到的成功数差异很大的可能样本与问题无关。在 Fisher(1945)看来，"……这些包含较少信息的可能性的存在不应当影响我们根据实际观察到的情况对显著性所做的判断……如此没有意义的结果可能发生的事实……在它们没发生的情况下绝对不应成为影响我们对显著性进行判断的理由；……只有相同类型的样本的抽样分布才有助于进行合理的显著性检验。"最近，Sprott(2000，第 6.4.4 节)提出了类似的观点。

Berger 和 Boos(1994)对无条件方法进行的调整在一定程度上回应了上述批评。他们在冗余参数 π 值的某一置信区间，而不是所有可能取值中求 P 值的上确界。其无条件检验的 P 值可表示为

$$P = \sup_{\pi \in C_\gamma} P_\pi(T \geqslant t_o) + \gamma,$$

其中 C_γ 是 π 的 $100(1-\gamma)\%$ 的置信区间。这里，γ 的取值非常小(如 0.001)，该检验能够确保置信水平的上界不变。

其他支持以两组边际总计为条件进行检验的论断指出，在很多问题上条件方法都提供了一种消除冗余参数的简单方法(例如，扩展到其他列联表问题)，并且边际本身几乎不包含任何关联的信息(Haber，1989；Yates，1984)。Zhu 和 Reid(1994)指出，除非 $\theta=1$，以边际总计为条件会损失一些信息。反对条件方法的论断部分担心由此导致的更严重的离散性问题。n_{11} 的可能取值很少，获得一个较小的 P 值非常困难。在使用一个名义置信水平进行重复检验时，实际发生第 I 类错误的概率可能比名义水平低得多，从而影响检验的统计效能。最后，关于非零值的统计推断(如置信区间)，我们将看到条件方法只适用于发生比之比，而对其他度量指标并不适用。

检验结果的保守性问题在一定程度上是无法避免的。离散分布的统计量在获得名义置信水平时必然会出现保守问题。相比之下，因为无条件检验只给定一个边际，其样本分布中可选表格的集合要大得多，因而，这个分布的离散性问题较轻，它所对应的 P 值的可能范围也比 Fisher 精确检验更为丰富。这样，通常情况下，无条件检验比 Fisher 精确检验的保守程度更低、统计效能更高。但它的缺点在于，对于较复杂的问题，如较大的表格，无条件检验的运算强度非常大。

当表格中确实包括两个独立的二项分布样本时，无条件检验似乎更为合理。关于这一点的讨论，请参见：Kempthorne(1979)。在其他情况下，条件检验更有意义。例如，在一项随机临床试验中，总数为 n 的便利样本(convenience sample)被随机分配到两个干预组。该样本不服从二项分布，因为它们不是所研究的两个总体的随机样本。我们可以只关注这个样本本身，考虑如果确实不存在干预效应的话，能观察到至少与现有结果一样极端的概率。例如，从 n 个对象中选取 n_{1+} 作为第 1 个干预组的所有可能方式中，有多大比例的情况 n_{11} 至少与所观察到的值一样大？在不存在干预效应的零假设下，无论怎样分配对象，总的成功和失败的结果分布(n_{+1}，n_{+2})应当是相同的。因此，列边际是自然给定的。与此相对应的是超几何零分布以及 Fisher 精确检验(Greenland，1981)。但是，这一结论不能扩展到关于非零效应的假设检验和置信区间。

当两组边际总计都自然给定时(如表 3.8)，不可避免地会出现高度离散问题，这时

Fisher 精确检验是最好的选择。此外，不论哪一组边际是自然给定的，使用中位 P 值都有助于降低离散性所带来的检验结果的保守问题。

3.5.7　精确条件分布的推导*

现在，我们展示独立性的条件检验如何服从超几何分布。由于接下来将要讨论对 Fisher 精确检验的扩展，在这儿，我们以 $I \times J$ 表格的情况为例。假定 $I \times J$ 表格中每一行的抽样都是独立的多项分布抽样，正如实际应用中常见的对 I 个干预组的比较。这时，行总计 $\{n_{i+}\}$ 是给定的，我们估计 I 个条件分布 $\{\pi_{j/i}, j=1,\cdots,J\}$。在"$H_0$：独立性"下，对于 $j=1,\cdots J$，存在 $\pi_{j/1}=\pi_{j/2}=\cdots=\pi_{j/I}=\pi_{+j}$。因而，这 I 个多项分布概率函数的乘积简化为

$$\prod_i \left(\frac{n_{i+}!}{\prod_j n_{ij}!} \prod_j \pi_{j/i}^{n_{ij}} \right) = \frac{(\prod_i n_{i+}!)(\prod_j \pi_{+j}^{n_{+j}})}{\prod_i \prod_j n_{ij}!}。 \tag{3.17}$$

这时，关于 $\{n_{ij}\}$ 的分布取决于 $\{\pi_{+j}\}$。这些是冗余参数，因为它们并不描述关联关系。Fisher 提出了消除冗余参数的一般方法，即以它们的充分统计量（sufficient statistics）为条件。从关于充分的定义可知，最后的条件分布不再取决于这些参数。

$\{\pi_{+j}\}$ 只有通过 $\{n_{+j}\}$ 才会影响公式 3.17 中多项分布概率的乘积，也即后者是前者的充分统计量。$\{n_{+j}\}$ 服从参数为 $(n, \{\pi_{+j}\})$ 的多项分布，即

$$\frac{n!}{\prod_j n_{+j}!} \prod_j \pi_{+j}^{n_{+j}}。 \tag{3.18}$$

由于 $\{n_{ij}\}$ 决定了 $\{n_{+j}\}$，$\{n_{ij}\}$ 和 $\{n_{+j}\}$ 的联合概率函数与 $\{n_{ij}\}$ 的概率函数完全相同。所以，以 $\{n_{+j}\}$ 为条件的 $\{n_{ij}\}$ 的概率函数，等于 $\{n_{ij}\}$ 的概率函数（公式 3.17）除以概率函数（公式3.18）在 $\{n_{+j}\}$ 处的取值，即

$$\frac{(\prod_i n_{i+}!)(\prod_j n_{+j}!)}{n! \prod_i \prod_j n_{ij}!}。 \tag{3.19}$$

这就是*多项超几何*（*multiple hypergeometric*）分布。它适用于与所观察到的表格具有相同的 $\{n_{i+}\}$ 和 $\{n_{+j}\}$ 的一组 $\{n_{ij}\}$。对于 2×2 表格，这就是公式 3.16 的超几何分布。

如果一个表格由单一的多项分布样本形成，未知参数为 $\{\pi_{ij}\}$。在进行独立性检验（对所有 i 和 j，$\pi_{ij}=\pi_{i+}\pi_{+j}$）时，公式 3.19 的分布是以行总计和列总计为条件的。这些边际总计是 $\{\pi_{i+}\}$ 和 $\{\pi_{+j}\}$ 的充分统计量，它们决定了零分布。无论在哪种抽样模型中，在以充分统计量为条件后两组边际都是给定的。最终，公式 3.19 不依赖于未知参数，因而可以计算精确的概率。

3.5.8　关于 $I\times J$ 表格独立性的精确检验*

对 $I\times J$ 表格的精确检验，我们使用多项超几何分布。Freeman 和 Halton（1951）将 P 值定义为在给定边际的可能表格中出现比观察到的表格更不可能发生的一组表格的概率。其他的精确检验将表格按照描述它与 H_0 之间距离的统计量进行排序，如 Yates（1934）使用了 X^2。这时，对于所观察到的表格对应的值 X_o^2，P 值就等于 $P(X^2 \geq X_o^2)$ 的零分布概率值。当分类变量是定序变量时，使用定序的统计量更为恰当。对于以存在正向关联为备择假设的情况，我们可以使用 $P(T \geq t_o)$，其中 T 可以是相关系数或者 γ，而 t_o 表示相应的观察值。

表 3.9 取自一项关于年轻妇女样本的吸烟量与心肌梗塞的个案-控制研究，我们通过

该表来展示有关定序变量的精确检验。由表 3.9 可见,第二行中的计数很小,因而不宜使用大样本检验。在给定边际计数的情况下,唯一一个更能反映吸烟与心肌梗塞正向关联的表格是第 1 行的计数为(25,26,11)、第 2 行的计数为(0,0,4)。以两组边际为条件,观察到的表格以及更极端的情况发生的零分布概率(按照公式 3.19)等于 0.018。尽管这一样本仅包含四位心肌梗塞病人,数据给出的证据显示,吸烟与心肌梗塞存在正向关联。该结果比使用 X^2 得到的结果更强,因为 X^2 并无考虑类别之间的排序。精确检验中,$P(X^2 \geqslant X_o^2) = P(X^2 \geqslant 6.96) = 0.052$。

表 3.9 精确条件检验的例子

	吸烟量(支/每天)		
	0	1 ~ 24	>25
控制组	25	25	12
心肌梗塞	0	1	3

来源:授权重印,基于:Table 5 in S. Shapiro et al. ,*Lancet* 743-746(1979)。

对 $I \times J$ 表格进行精确检验的算法和软件均很常见(如:Mehta and Patel, 1983;另见附录 A)。当渐近近似方法不适用时,我们建议使用这些精确检验方法。精确检验的运算时间随着 n,I 或 J 的增加而呈指数上升。然而,运用蒙特卡洛法(Monte Carlo)可以在给定边际总计的情况下随机选取一组表格。这时,估计的 P 值等于具有至少与检验统计量的观察值一样大的表格所占的样本比例。

随着 I 和/或 J 的增加,任何检验统计量 T 的可能取值数往往也会增加。因此,条件检验所受保守性问题的影响会降低。

3.6 2×2 表格的小样本置信区间*

小样本方法同样也适用于统计估计。可以遵循前面的思路得出仅取决于所关注的参数的精确分布。这些分布是估计诸如发生比之比等度量指标的置信区间的基础。

3.6.1 发生比之比的小样本统计推断

在多项分布抽样的情况下,$\{n_{ij}\}$ 的分布取决于 n 和单元格概率 $\{\pi_{ij}\}$。对于 2×2 表格,发生比之比为

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\pi_{11}(1 - \pi_{1+} - \pi_{+1} + \pi_{11})}{(\pi_{1+} - \pi_{11})(\pi_{+1} - \pi_{11})}。$$

因此,π_{11} 是一个关于 θ 和 $\{\pi_{1+}, \pi_{+1}\}$ 的函数。同样的推导适用于所有的 π_{ij},所以 $\{n_{ij}\}$ 的多项分布可以使用参数 $\{\theta, \pi_{1+}, \pi_{+1}\}$ 来表示。在 $\{n_{1+}, n_{+1}\}$ 给定的条件下,$\{n_{ij}\}$ 的分布仅取决于 θ。由于在给定边际总计后,n_{11} 决定了其他所有的单元格计数,$\{n_{ij}\}$ 的条件分布可以通过函数 $P(n_{11} = t) = f(t; n_{1+}, n_{+1}, n, \theta)$ 来表示。这个分布(Fisher, 1935c)是*非中心超几何*(*non-central hypergeometric*)分布,

$$f(t; n_{1+}, n_{+1}, n, \theta) = \frac{\binom{n_{1+}}{t}\binom{n - n_{1+}}{n_{+1} - t}\theta^t}{\sum_{u=m_-}^{m_+}\binom{n_{1+}}{u}\binom{n - n_{1+}}{n_{+1} - u}\theta^u} \tag{3.20}$$

其中 $m_- \leqslant t \leqslant m_+$。

在观察到的 $n_{11}=t_o$ 时,对检验 $H_0: \theta=\theta_0$ 进行逆向运算便可得出 θ 的置信区间。对于 $H_a: \theta>\theta_0$,检验的 P 值为

$$P = \sum_{t \geqslant t_o} f(t; n_{1+}, n_{+1}, n, \theta_0)。$$

对于 $H_a: \theta<\theta_0$,

$$P = \sum_{t \leqslant t_o} f(t; n_{1+}, n_{+1}, n, \theta_0)。$$

当 $\theta_0=1$ 时,以上检验就是单边的 Fisher 精确检验。Cornfield(1956)使用*尾部法*(*tail method*)构建了 θ 的置信区间,其下限是在备择假设为 $H_a: \theta>\theta_0$ 的检验中 $P=\alpha/2$ 对应的 θ_0,上限是在备择假设为 $H_a: \theta<\theta_0$ 的检验中 $P=\alpha/2$ 对应的 θ_0。这个区间由一组满足两个单边 P 值 $\geqslant\alpha/2$ 的 θ_0 值构成。

与 Fisher 精确检验一样,通过条件方法得出的区间估计必然会因离散性问题而偏于保守。实际的置信水平,定义为对所有可能 θ 值的涵盖概率的下确界(infimum),以名义置信水平为其下限。由一个双边检验求逆所形成的区间,比对两个单边检验求逆所得结果保守程度较轻且区间较短(Agresti and Min, 2001;Baptista and Pike, 1977)。在独立的二项分布样本中,可供选择的另一种办法是对非零的无条件小样本检验求逆。由于这一方法降低了离散性问题的影响,因而所得区间也往往较短。

关于 θ 的*条件最大似然估计*(*conditional ML estimate*)是使概率(公式 3.20)最大化的 θ 值。由相应的对数似然函数对 θ 求导可知,该估计值满足关于 θ 的等式 $n_{11}=E(n_{11})$,其中期望值是针对公式 3.20 分布而言的。这个等式具有唯一解 $\hat{\theta}$,并可通过迭代法求得(Cornfield, 1956)。该估计值不同于*无条件最大似然估计值*(*unconditional ML estimator*)$\hat{\theta}=n_{11}n_{22}/n_{12}n_{21}$,后者利用了 $\{n_{ij}\}$ 的多项分布中关于 $\{\pi_{ij}\}$ 的最大似然估计。利用统计软件,我们可以计算关于发生比之比的条件最大似然估计及其小样本置信区间(如相应的 SAS 程序,参见表 A.2)。

3.6.2 例子:品茶实验

我们以表 3.8 中 Fisher 的品茶实验为例来加以说明。对 θ 的条件最大似然估计等于 6.4。软件报告的 Cornfield 尾部法区间为(0.2,626.2),置信水平确保 $\geqslant$0.95。这个结果并不令人惊讶,由于样本太小,所以置信区间非常宽。由双边“精确”条件计分检验求逆所得出的区间估计更为精确,即为(0.3,306.2)。考虑到数据抽样设计的特点,无条件方法在此并不适用(假设该表包括的是两个二项分布样本,通过对无条件“精确”计分检验求逆所得的区间为(0.4,234.4))。

3.6.3 离散性对精确置信区间的影响

就条件分布不受冗余参数影响的事实而言,小样本统计推断是“精确的”。相应的置信区间和假设检验运用了精确的概率计算而不是近似分布。然而,由于离散性的存在,它们在操作上具有保守性的特点。

大样本方法并不一定能保证误差概率的边界,其结果既可能偏保守,也可能过于宽松,因而大样本方法与精确方法的结果可能并不一致。例如,对于品茶实验数据(表 3.8),皮尔逊卡方检验的 P 值等于 0.157,与之相比,双边精确检验的 P 值为 0.486。关

于发生比之比的95%的大样本置信区间(公式3.2)为(0.4,220.9),而相应的Cornfield精确区间为(0.2,626.2)。一般来说,我们可能会偏好精确方法而不是近似方法。然而,当条件分布高度离散时,到底应当使用哪种方法并不是很明确。这时,精确方法给出的结果非常保守,尤其是在小样本的情况下。

对于高度离散的数据,使用中位P值对精确方法进行调整似乎更为合理。这样,条件方法下的置信区间是利用中位P值对超几何检验$\theta=\theta_0$求逆而得出的。尽管无法保证误差的概率不超过名义水平,但它通常会比精确方法给出的结果更接近于理想的水平。与大样本方法相比,它具有随着离散性的降低而表现更好的优点,因为这种情况下它实际上与使用普通的P值进行相应的精确检验是一样的。

基于中位P值的统计推断是对精确方法的保守性与大样本方法的不确定性的一种折中。对于发生比之比的区间估计,这种方法所得结果一般略偏保守,但是在小样本情况下,它会给出比Cornfield精确区间小得多的区间。例如,就品茶数据而言,使用中位P值对两个单边超几何检验求逆所得的95%的置信区间为(0.31,309),与之相比,Cornfield区间为(0.2,626.2)。

3.6.4 比例之差的小样本统计推断

消除冗余参数的条件方法适用于那些具有充分统计量的参数。然而,后面我们将看到(第6.7.9节),简化的充分统计量仅仅在某些模型下才会存在。对于二分数据,这样的模型必须以发生比之比为参数。就2×2表格来说,通过条件方法无法得出比例之差或比例之比的置信区间。无条件方法更为复杂,但它不需要充分统计量。在第3.5.5节中,我们已使用该方法对独立二项样本进行了关于$\pi_1-\pi_2=0$的检验。

小样本置信区间是通过对任意给定的$-1<\delta_0<1$,将无条件检验H_0: $\pi_1-\pi_2=\delta_0$进行求逆而得的。表格的概率函数等于$\text{bin}(n_1,\pi_1)$和$\text{bin}(n_2,\pi_2)$两个密度函数的乘积。该函数可由$\delta=\pi_1-\pi_2$和冗余参数λ来表示。例如,如果$\lambda=\pi_1+\pi_2$,我们可以代入$\pi_1=(\lambda+\delta)/2$和$\pi_2=(\lambda-\delta)/2$。对于$\delta=\delta_0$以及给定的λ值,我们可以使用二项分布乘积来计算检验统计量至少与其观察值一样大的概率。P值是针对λ的所有可能取值对应概率的上确界。这给出了关于不同δ_0值的一组检验。$\pi_1-\pi_2$的置信区间就是使得这个P值大于α的一组δ_0。

上述这种方法得出的结果可能会非常保守。有关各种检验统计量的详细讨论,参见:Agresti and Min(2001)、Coe and Tamhane(1993)、Santner and Snell(1980)、Santner and Yamagami(1993)。如Coe and Tamhane(1993)所述,对一个双边检验求逆所得结果一般要好于对两个单边检验求逆。

3.7 对多维表格以及非表格形式结果变量的扩展

本章所介绍的方法可以扩展到多维列联表。例如,对两维表格的独立性检验可以扩展到三维表格中的条件独立性检验。在后面的章节中,我们将通过在模型中界定有关参数及其统计推断来介绍相应扩展。这些方法可以应用到更一般的情况,比如当某些解释变量是连续而不是分类变量的情况。

3.7.1 分类数据不一定表示为列联表的形式

本章前面的例子所介绍的分类数据都对应着列联表的形式。然而,本书所关注的内

容远比列联表分析更为广泛。有关分类结果变量的模型既可以包括连续的,也可以包括分类的解释变量,甚至当所有或大部分变量都是分类变量时,数据的源文件通常也不是列联表的形式,而是每个对象对应着数据中的一行。一个包括调查对象的性别、种族、受教育程度(1 = 高中以下,2 = 高中或大学未毕业,3 = 大学毕业)以及对同性恋的看法(1 = 宽容,2 = 憎恶)的数据文件,其前三行可能为:

对象	性别	种族	教育	看法
1	f	w	2	1
2	m	b	3	1
3	m	w	1	2

软件可以读取这种格式的数据文件,并进行包括构建列联表在内的各种分析。

在下一章,我们将介绍本书余下部分所使用的模型框架。我们在本章中已讨论的所有方法均来自于对简化形式的模型中参数的统计推断。

注 解

第 3.1 节:关联参数的置信区间

3.1 利用 $\log\theta$ 的 Woolf 区间(式 3.2)来处理零单元格计数问题的研究包括:Agresti(1999)、Gart(1966,1971)。Goodman(1964a)同时给出了 $I \times J$ 表格中的所有发生比之比的区间。Brown 和 Benedetti(1977)以及 Goodman 和 Kruskal(1963,1972)介绍了许多关联度量指标的标准误。Goodman 和 Kruskal(1963,1972)将公式 3.9 扩展到了独立的多项分布抽样的情况。

3.2 Agresti 和 Caffo(2000)表明,与在单一样本情况下相同(习题 1.24),关于 $\pi_1 - \pi_2$ 的沃尔德区间(式 3.4)在对每个类别加上两个虚拟观测值(pseudo-observations)(每个样本中的每一类别各一个)后表现得明显更好。

第 3.2 节:二维列联表的独立性检验

3.3 对于超几何分布样本,在独立性检验中 $\{\hat{\mu}_{ij}\}$ 是*精确的*(而不是*估计的*)期望值。具体地,

$$E(n_{11}) = \frac{n_{1+}n_{+1}}{n} \quad \text{且} \quad \text{var}(n_{11}) = \frac{n_{1+}n_{+1}n_{2+}n_{+2}}{n^2(n-1)}。$$

Haldane(Haldane, 1940)推导出 $E(X^2) = (I-1)(J-1)n/(n-1)$ 以及关于计算 $\text{var}(X^2)$ 的一个复杂公式;Dawson(Dawson, 1954)给出了相应的简化表述。Lewis 等(Lewis et al., 1984)推导了第三阶中心矩量。Watson(1959)证明了 X^2 的条件分布也服从极限卡方分布。

3.4 Diaconis 和 Efron(1985)介绍了具有相同的 I, J 和 n 的所有可能表格服从均匀分布的统计推断;他们的*大量数据检验*(*volume test*)考虑满足 $X^2 \leqslant X_o^2$ 的表格所占的比例。

3.5 对于复杂的抽样设计,有必要使用专门的方法。在生物医学应用中,经常使用序贯法(sequential methods)(Jennison and Turnbull, 2000, Chap. 12)。在社会科学应用中,经常涉及整群和/或分层抽样。La Vange 等(La Vange et al., 2001)以及 Rao 和 Thomas(1988)综述了复杂抽样方法下的分类数据分析。Gleser 和 Moore(1985)指出,正相关会导致皮尔逊统计量的零分布随机上升。另见:Bedrick(1983)、Clogg and Eliason(1987)、Fay(1985)、Holt et al.(1980), Koehler and Wilson(1986)、Rao

and Scott(1987)、Scott and Wild(2001)、Shuster and Downing(1976)、Tavare and Altham(1983),以及本书第12章介绍的方法。

当某些数据缺失时,有必要进行一定的修正。Watson(1956)可能是第一个对此进行研究的统计学家。在假定无应答可以忽略的情况下,Lipsitz 和 Fitzmaurice(1996)推导了关于列联表的独立性和条件独立性的计分检验,他们证明检验统计量渐近服从常见的卡方零分布。对这些方法的综述,参见 Schafer(1997,Chap.7)。

第3.4节:定序变量的二维表格

3.6 Bhapkar(1968)以及 Yates(1948)提出了与 M^2 相似的统计量,并介绍了有关单一排序表格的统计量。Graubard 和 Korn(1987)列出了14种对 $2\times J$ 表格利用某种相关系数统计量进行的检验。另见:Nair(1987)、Williams(1952)。Cohen 和 Sackrowitz(1991,1992)进行了决策论方面的评估,如基于 γ 以及局部对数发生比之比的检验的可容许性(admissibility)。Rayner 和 Best(2001)考虑了有关列联表分析的非参数方法。

第3.5节:小样本的独立性检验

3.7 Yates(1934)提到 Fisher 向他建议使用超几何分布来进行精确检验。他提出了一个经过连续性修正后的 X^2,即

$$X_c^2 = \sum\sum \frac{(|n_{ij} - \hat{\mu}_{ij}| - 0.5)^2}{\hat{\mu}_{ij}},$$

以此来近似精确检验。Haber(1980,1982)、Plackett(1964),以及 Yates(1984)讨论了它的合理性。基于现有软件,即使对大样本进行 Fisher 精确检验也是可行的,因而也就不再需要这个修正。

3.8 Fisher 精确检验的 UMPU 特性来自于以一个完全的、服从指数族分布的充分统计量为条件(Lehmann, 1986,Secs 4.5-4.7)。Fleiss(1981)、Gail 和 Gart(1973),以及 Suissa 和 Shuster(1985)研究了在 Fisher 检验中为了达到固定的统计效能所需的样本规模。关于条件化的争论包括:Barnard(1945, 1947, 1949, 1979), Berkson(1978),Fisher(1956), Howard(1998), Kempthorne(1979), Lloyd(1988a), Pearson(1947), Rice(1988), Routledge(1992), Suissa and Shuster(1984, 1985), Yates(1984)。Yates 与评述者还讨论了有关选取双边 P 值的问题。对无条件方法的讨论包括:Chan(1998), Martin Andres and Silva Mato(1994), Rohmel and Mansmann(1999)。Altham(1969)以及 Howard(1998)讨论了关于 2×2 表格的贝叶斯分析(参见第15.2.3节)。Agresti(1992,2001)对小样本方法进行了综述。

3.9 对于利用中位 P 值进行统计推断的讨论,参见:Berry and Armitage(1995)、Hirji(1991)、Hwang and Wells(2002)、Hwang and Yang(2001)、Mehta and Walsh(1992)、Routledge(1994)。也可以通过其他形式的 P 值进行类似的改进。一种方法是,利用表格概率来生成一个对样本空间的更精细的分割,在当几个表格的检验统计量的取值相同时,这种方法很有用;对于具有与观察到的检验统计量相同值的表格,只有那些发生可能性不超过所观察到的表格的表格才影响 P 值(Cohen and Sackrowitz, 1992;Kim and Agresti, 1995)。这不仅仅取决于充分统计量,在某些情况下它的 Rao-Blackwellized 形式就是中位 P 值(Hwang and Wells, 2002)。通过未对离散性问题进行连续性修正的高阶渐近方法得到的普通 P 值具有与中位 P 值相似的效果(Pierce and Peters, 1999;Strawderman and Wells, 1998)。

3.10 有关 $I \times J$ 表格的精确分析，参见：Mehta and Patel(1983)。对于定序变量的情况，另见：Agresti et al. (1990)。有关精确 P 值的蒙特卡洛估计，参见：Agresti et al. (1979)、Booth and Butler(1999)、Diaconis and Sturmfels(1998)、Forster et al. (1996)、Mehta et al. (1988)、Patefield(1982)。Gail 和 Mantel(1977)以及 Good(1976)给出了计算满足某种给定边际的表格数的近似公式。Freidlin 和 Gastwirth(1999)扩展了无条件方法，用以对 $I \times 2$ 表格进行趋势检验以及对几个 2×2 表格进行条件独立性检验。

第 3.6 节：2×2 表格的小样本置信区间

3.11 假定 (θ, λ) 存在最小充分统计量 (T, U)，其中 λ 是一个冗余参数，如果 U 的分布只取决于 λ，并且在给定 U 的条件下 T 的分布仅取决于 θ，Cox 和 Hinkley(1974, p. 35)将 U 定义为*附属*(*ancillary*)于 θ。对于发生比之比为 θ 和 $\lambda = (\pi_{1+}, \pi_{+1})$ 的 2×2 表格，令 $T = n_{11}$ 以及 $U = (n_{1+}, n_{+1})$，那么 U 不是附属的，因为它的分布既取决于 θ，又取决于 λ。利用 Godambe 的定义，Bhapkar(1989)称边际计数 U *部分附属*(*partial ancillary*)于 θ。这意味着，在给定 U 的情况下，数据的分布仅取决于 θ，并且在给定 θ 的情况下，对于不同的 λ 来说 U 的分布族是完整的。Liang(1984)给出了另一个关于条件和无条件统计推断同样有效的定义。

习题

应用部分

3.1 参见表 2.9，构建总体中关于以下参数的 95% 置信区间并加以解释：(a)发生比之比，(b)比例之差，(c)相对风险。

3.2 参见表 2.5 中的有关肺癌与吸烟的数据，构建一个针对相应的关联度量指标的置信区间，并予以解释。

3.3 在 1980—1982 年的职业篮球比赛中，当波士顿凯尔特人队的拉里·伯德(Larry Bird)进行两个罚篮时，有 5 次两罚全失，251 次两罚全中，34 次只罚中第一个，并且 48 次只罚中第二个(Wardrop, 1995)，有没有可能两次连续罚篮之间是独立的？

3.4 参见表 3.10。

a. 利用 X^2 和 G^2，对政党认同与种族进行独立性检验，报告相应的 P 值并加以解释。

b. 通过残差来描述二者之间存在关联的证据。

c. 将卡方分割为两项，一项是针对在民主党和独立党派之间的选择，另一项是针对将前两者合并后与共和党之间的选择，加以解释。

d. 构建一个关于种族与是否为民主党或共和党的发生比之比的 95% 置信区间来对关联进行总结，加以解释。

表 3.10 习题 3.4 的数据

	政党认同		
种族	民主党	独立党派	共和党
黑人	103	15	11
白人	341	105	405

来源：1991 年综合社会调查(General Social Survey)，美国民意研究中心(National Opinion Research Center)。

3.5 参见表 3.10。根据同一个调查数据，对性别和政党认同进行了列联表分析。表 3.11 给出了一些结果。说明应该怎么解释这些输出结果。

表3.11 习题3.5的结果

Frequency Expected	dem	indep	repub
female	279 261.42	73 70.653	225 244.93
male	165 182.58	47 49.347	191 171.07

Statistic	DF	Value	Prob
Chi-Square	2	7.0095	0.0301
Likelihood Ratio Chi-Square	2	7.0026	0.0302

Observ	Resraw	Reschi	StReschi	Observ	Resraw	Reschi	StReschi
1	17.584	1.088	2.293	4	−17.584	−1.301	−2.293
2	2.347	0.279	0.465	5	−2.347	−0.334	−0.464
3	−19.931	−1.274	−2.618	6	19.931	1.524	2.618

译者注——female:女性;male:男性;dem:民主党;indep:独立党派;repub:共和党。

3.6 在一项关于在就诊时乳腺癌的发展阶段(局部或晚期)与妇女的居住方式之间的关系的研究中:在144个独居的妇女中,41.0%为晚期;在209名与配偶同住的妇女中,52.2%为晚期;在89名与其他人同住的妇女中,59.6%为晚期。研究者报告说它们之间的关系的 P 值为0.02(D. J. Moritz and W. A. Satariano, *J. Clin. Epidemiol.* 46: 443-454, 1993)。重新分析这些数据以求出这个 P 值。

3.7 参见表2.1,将对心脏病发作是否独立于服用阿司匹林的检验的 G^2 分割为两项,加以解释。

3.8 《项目蓝皮书:不明航空器分析报告(*Project Blue Book: Analysis of Reports of Unidentified Aerial Objects*)》是由美国空军在1955年5月出版的(莱特-帕特森空军基地空中技术智能中心,Air Technical Intelligence Center at Wright-Patterson Air Force Base)来分析有关不明飞行物(unidentified flying objects, UFOs)的报告。在它的表Ⅱ中,该报告按照物体的颜色(九种)进行了划分,将1 765次观测视为已知物体,434次观测视为不明物体。该报告称,"卡方检验仅仅适用于具有相同数量的元素的分布",因此研究者在计算 X^2 前将已知类别中的每个计数都乘以(434/1 765),使每一行都有434个观测值。他们报告说 $X^2=26.15$,自由度为df=8。解释为什么这是错误的。X^2 应当等于多少?(提示:对于他们调整后的表格,首先证明在一列中每个单元格对 X^2 的贡献是一样的,接着证明将一行中的每个单元格计数都乘以一个常数对此的影响)

3.9 表3.12对一个精神病患者的样本按照诊断结果与是否要求药物治疗进行了划分。
a. 求出关于独立性的标准化皮尔逊残差,并加以解释。
b. 将卡方分割为三项来描述诊断结果间的异同,比较:(i)前两行,(ii)第三和第四行,(iii)最后一行与前两行的合并以及第三、四行的合并。

3.10 参见表7.8。对于将两个性别的数据进行合并所得到一个4×4表格,$X^2=11.5$ ($P=0.24$),然而在使用行赋值为(3,10,20,35)以及列赋值为(1,3,4,5)时,$M^2=7.04$ ($P=0.008$)。解释为什么结果会有这么大的差别。

表 3.12 习题 3.9 的数据

诊断结果	药物治疗	非药物治疗
精神分裂症	105	8
情感障碍	12	2
神经官能症	18	19
人格障碍	47	52
特殊症状	0	13

来源：授权重印自：E. Helmes and G. C. Fekken, *J. Clin. Psychol.* 42: 569-576(1986)。

3.11 一项关于高中学生的教育抱负的研究(S. Crysdale, *Internat. J. Compar. Sociol.* 16: 19-36, 1975)用以下尺度来测量抱负(未完成高中、高中毕业、大学未毕业、大学毕业)。当家庭收入为低收入时，在这些类别中的学生计数为(11,52,23,22)；当家庭收入为中等收入时的计数为(9,44,13,10)；以及当家庭收入为高收入时的计数为(9,41,12,27)。

a. 通过 X^2 或 G^2 检验教育抱负与家庭收入的独立性，说明对该数据进行这种检验的局限性。

b. 求出标准化皮尔逊残差。这些残差有没有表明任何关联模式？

c. 进行一个统计效能更高的检验，加以解释。

3.12 参见表 8.15，求一个关于 γ 的 95% 置信区间，解释受教育程度与对流产的态度之间的关联。

3.13 表 3.13 给出的是一项关于比较通过放射疗法和手术治疗喉癌的回顾性研究。被访者回答是否至少在治疗后的两年内癌症得到了控制。表 3.14 给出了 SAS 的输出结果。

a. 报告并解释以下 Fisher 精确检验的 P 值：(i) $H_a: \theta>1$，(ii) $H_a: \theta\neq 1$。说明 P 值是怎样计算出来的？

b. 解释所有关于 θ 的置信区间，说明它们之间的差异以及它们是怎样计算出来的。

c. 找出并解释单边的中位 P 值，说出这种 P 值的优点和缺点。

表 3.13 习题 3.13 的数据

	癌症得到控制	癌症未得到控制
手术	21	2
放射疗法	15	3

来源：授权重印自：W. M. Mendenhall, R. R. Million, D. E. Sharkey, and N. J. Cassisi, *Internat. J. Radiat. Oncol. Biol. Phys.* 10: 357-363(1984), Pergamon Press Plc.

3.14 一项研究考察在患有转移性乳腺癌的妇女中脱氢皮质醇对高钙血的影响(B. Kristensen et al., *J. Intern. Med.* 232: 237-245, 1992)。在 30 个病人中，15 人被随机选出使用脱氢皮质醇，另外 15 人构成了控制组。干预组中有 7 个病人的免疫血清离子化钙达到了正常水平，控制组中一个也没有。分析是否这项治疗具有显著的效果，加以解释。

3.15 对于习题 3.14，使用以下方法得出关于发生比之比的 95% 置信区间：(a) Woolf(即沃尔德)区间，(b) Cornfield“精确”方法，(c) 剖面似然函数。在每一种情况下，指出零单元格计数的影响。综述每种方法的优缺点。

表 3.14　习题 3.13 的 SAS 输出结果

Fisher's Exact Test	
Cell (1,1) Frequency (F)	21
Left-sided Pr <= F	0.894 7
Right-sided Pr >= F	0.380 8
Table Probability (P)	0.275 5
Two-sided Pr<= P	0.638 4
Odds Ratio	2.100 0

Asymptotic Conf Limits:	95% Lower Conr Limit	0.311 6
	95% Upper Conf Limit	14.152 3
Exact Conf Limits:	95% Lower Conf Limit	0.208 9
	95% Upper Conf Limit	27.552 2

3.16　参考品茶数据(表 3.8)。利用普通的 P 值和中位 P 值,分别构建关于 $H_a: \theta > 1$ 的 Fisher 精确检验的零分布,求出它们的期望值并进行比较。

3.17　考虑一个具有以下数值的 3×3 表格,按行分别为(4,2,0/2,2,2/0,2,4)。利用 X^2 进行一个独立性精确检验。假定行和列是定序的,使用等距的赋值进行一个定序精确检验,说明为什么结果差别很大。

3.18　在 1999 年先灵公司(Schering Corp.)关于过敏药的一个广告中提及,在一个儿科的随机临床试验中,使用氯雷他定(loratadine)克敏能(Claritin)的 188 名患者中 4 人出现神经过敏的症状,262 名使用安慰剂的病人中 2 人出现该症状,170 名使用氯苯胮敏(choropheniramine)的病人中 2 人出现该症状。在以下每一部分,说明你使用的是哪种方法以及为什么。

a. 统计推断结果能表明神经过敏和药物有关吗?

b. 对于服用克敏能和安慰剂的组,构建以下的 95% 置信区间并加以解释:(i)发生比之比,(ii)出现神经过敏的比例之差。

3.19　参考习题 2.19 中关于性生活乐趣的数据,对这些数据进行分析,写一个简短报告来总结分析结果并加以解释。

理论与方法

3.20　$\hat{\theta}$ 是有关 θ 的大样本和小样本置信区间的中点吗? 为什么是或为什么不是?

3.21　对于比较两个二项分布样本,证明对数发生比之比的标准误(式 3.1)随着给定样本中的成功和失败所占的比例之差的增加而增加。

3.22　通过 δ 方法,证明关于二项分布参数 π 的 Logit 的沃尔德置信区间是

$$\log[\hat{\pi}/(1-\hat{\pi})] \pm z_{\alpha/2}/\sqrt{n\hat{\pi}(1-\hat{\pi})}。$$

说明如何利用这个区间获得一个关于 π 本身的置信区间(Newcombe(2001)指出,在 Logit 刻度上,样本 Logit 也是关于 π 的计分区间的中点。他表明这个 Logit 区间包含了计分区间)。

3.23　对于两个参数,基于单一样本估计 $\hat{\theta}_i$ 和区间(l_i, u_i),其中 $i=1,2$,$\theta_1-\theta_2$ 的一个置信区间为

$$\left(\hat{\theta}_1-\hat{\theta}_2-\sqrt{(\hat{\theta}_1-l_1)^2+(u_2-\hat{\theta}_2)^2},\quad \hat{\theta}_1-\hat{\theta}_2+\sqrt{(u_1-\hat{\theta}_1)^2+(\hat{\theta}_2-l_2)^2}\right)。$$

Newcombe(1998b)利用关于 π_i 的计分区间(l_i, u_i)提出了关于 $\pi_1 - \pi_2$ 的一个区间，它大大优于沃尔德区间(式 3.4)。该区间为$(\hat{\pi}_1 - \hat{\pi}_2 - z_{\alpha/2}s_L, \hat{\pi}_1 - \hat{\pi}_2 + z_{\alpha/2}s_U)$，其中

$$s_L = \sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}}, \quad s_U = \sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{l_2}}。$$

证明它具有以上关于 $\theta_1 - \theta_2$ 的区间的一般形式。

3.24 对于多项分布抽样，利用 $\log\hat{\theta}$ 的渐近方差证明，对于 Yule 的 Q(习题 3.26)，$\sqrt{n}(\hat{Q} - Q)$的渐近方差是 $\sigma^2 = (\sum_i \sum_j \pi_{ij}^{-1})(1 - Q^2)^2/4$(Yule, 1900, 1912)。

3.25 参考习题 2.23。针对多项分布抽样，证明如何通过首先找出一个关于 $\log(1 - AR)$ 的置信区间来得到关于 AR 的置信区间(Fleiss, 1981, p. 76)。

3.26 对于任意维数列联表中的多项分布概率 $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots)$，假定存在一个度量指标 $g(\boldsymbol{\pi}) = v/\delta$。证明 $\sqrt{n}[g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})]$的渐近方差等于 $\sigma^2 = [\sum_i \pi_i \eta_i^2 - (\sum_i \pi_i \eta_i)^2]/\delta^4$，其中 $\eta_i = \delta(\partial v/\partial \pi_i) - v(\partial \delta/\partial \pi_i)$(Goodman and Kruskal, 1972)。

3.27 对于定序变量，考虑 γ(式 2.14)。令

$$\pi_{ij}^{(c)} = \sum_{a<i}\sum_{b<j}\pi_{ab} + \sum_{a>i}\sum_{b>j}\pi_{ab}, \quad \pi_{ij}^{(d)} \sum_{a<i}\sum_{b>j}\pi_{ab} + \sum_{a>i}\sum_{b<j}\pi_{ab},$$

其中 i 和 j 在求和过程中是给定的。证明 $\prod_c = \sum_i \sum_j \pi_{ij}\pi_{ij}^{(c)}$ 以及 $\prod_d = \sum_i \sum_j \pi_{ij}\pi_{ij}^{(d)}$。通过 δ 方法证明 $\hat{\gamma}$ 服从大样本正态分布(式 3.9)，其中(Goodman and Kruskal, 1963)

$$\phi_{ij} = 4[\prod_d \pi_{ij}^{(c)} - \prod_c \pi_{ij}^{(d)}]/(\prod_c + \prod_d)^2, \quad \sum_i \sum_j \pi_{ij}\phi_{ij} = 0,$$

$$\sigma^2 = \frac{16}{(\prod_c + \prod_d)^4}\sum_i \sum_j \pi_{ij}[\prod_d \pi_{ij}^{(c)} - \prod_c \pi_{ij}^{(d)}]^2。$$

3.28 一个 $I \times J$ 表格的列是定序的而行不是定序的。Ridits(Bross, 1958)是基于数据对列的赋值。第 j 个样本 ridit 就是在第 j 类内的平均累积比例，

$$\hat{r}_j = \sum_{k=1}^{j-1} p_{+k} + \left(\frac{1}{2}\right)p_{+j}。$$

第 i 行的样本 ridit 的均值为 $\hat{R}_i = \sum_j \hat{r}_j p_{j|i}$。证明 $\sum_j p_{+j}\hat{r}_j = 0.50$ 并且 $\sum_i p_{i+}\hat{R}_i = 0.50$ (有关 ridit 分析，参见：Agresti(1984, Secs. 9.3 and 10.2), Bross(1958), Fleiss(1981, Sec. 9.4), Landis et al. (1978))。

3.29 证明 $X^2 = n\sum\sum(p_{ij} - p_{i+}p_{+j})^2/p_{i+}p_{+j}$。因此，在大样本情况下，无论关联是否实际存在，$X^2$ 的值都会很大。说明为什么这个检验与其他检验一样，只是表明了拒绝 H_0 的证据的程度，而没有描述关联的强度。("像火一样，卡方检验是个优秀的仆人，却是个糟糕的主人")(Sir Austin Bradford Hill, *Proc. Roy. Soc. Med.* 58: 295-300, 1965)。

3.30 对于利用在 n_1 和 n_2 次试验中的独立的二项分布变量 y_1 和 y_2 进行检验 $H_0: \pi_1 = \pi_2$，计分统计量为

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1 - \hat{\pi})(1/n_1 + 1/n_2)}},$$

其中 $\hat{\pi}=(y_1+y_2)/(n_1+n_2)$ 是在 H_0 成立时关于 $\pi_1=\pi_2$ 的联合估计(*pooled estimate*)。证明 $z^2=X^2$。

3.31 对于一个 2×2 表格,考虑 H_0: $\pi_{11}=\theta^2$, $\pi_{12}=\pi_{21}=\theta(1-\theta)$, $\pi_{22}=(1-\theta)^2$。

a. 证明它的边际分布满足独立同分布。

b. 对于一个多项分布样本,在 H_0 下证明 $\hat{\theta}=(p_{1+}+p_{+1})/2$。

c. 说明如何检验 H_0,证明检验统计量的自由度为 df=2。

d. 参见习题 3.3。拉里·伯德的两次罚球有可能满足独立同分布吗?

3.32 对于一个 2×2 表格,证明:

a. 四个皮尔逊残差可以取不同的值。

b. 所有四个标准化皮尔逊残差具有相同的绝对值(由于 df=1,这是合理的)。

c. 每个标准化皮尔逊残差的平方等于 X^2(注意:对于 2×2 表格,$X^2=n(n_{11}n_{22}-n_{12}n_{21})^2/(n_{1+}n_{2+}n_{+1}n_{+2})$。有关 $I\times J$ 表格的 X^2 公式,参见:Mirkin(2001))。

3.33 关于独立性检验,证明 $X^2\leq n\min(I-1,J-1)$,因而 $V^2=X^2/[n\min(I-1,J-1)]$ 的取值落在 0 和 1 之间(Cramér, 1946)。对于 2×2 表格,X^2/n 常常被称为 ϕ^2(*phi-squared*),它等于 Goodman 和 Kruskal 的 τ(习题 2.38)。其他的基于 X^2 的度量指标包括*列联系数*(*contingency coefficient*)$[X^2/(X^2+n)]^{1/2}$(皮尔逊 1904)。

3.34 对于计数 $\{n_i\}$,检验拟合优度的*效能多样化统计量*(*power divergence statistic*)(Cressie and Read, 1984;Read and Cressie, 1988)是:

$$\text{对于}\ -\infty<\lambda<\infty,\quad \frac{2}{\lambda(\lambda+1)}\sum n_i\left[(n_i/\hat{\mu}_i)^{\lambda}-1\right]。$$

a. 当 $\lambda=1$ 时,证明它等于 X^2。

b. 随着 $\lambda\to0$,证明它收敛于 G^2(提示:$\log t=\lim_{h\to0}(t^h-1)/h$)。

c. 随着 $\lambda\to-1$,证明它收敛于 $2\sum\hat{\mu}_i\log(\hat{\mu}_i/n_i)$,即*最小判别信息*(*minimum discrimination information*)统计量(Gokhale and Kullback, 1978)。

d. 当 $\lambda=-2$ 时,证明它等于 $\sum(n_i-\hat{\mu}_i)^2/n_i$,即*纽曼修正卡方*(*Neyman modified chi-squared*)统计量(Neyman, 1949)。

e. 当 $\lambda=-\frac{1}{2}$ 时,证明它等于 $4\sum\left(\sqrt{n_i}-\sqrt{\hat{\mu}_i}\right)^2$,即*弗里曼-涂盖*(*Freeman-Tukey*)统计量(Freeman and Tukey, 1950)。

(在正则性条件下,它们的渐近分布是一致的(参见 Drost et al., 1989)。当 λ 的值接近 $\frac{2}{3}$ 时,它对卡方零分布的近似最好。)

3.35 通过分割的思路来说明为什么独立性检验的 G^2 在将一个列联表的两行(或两列)合并后不可能增加(提示:证明完整表格的 G^2 = 合并后表格的 G^2 + 被合并的那两行所组成的表格的 G^2)。

3.36 通过证明关于 $\{n_{ij}\}$ 的多项超几何分布(式 3.19)可以分解为各自的表格所对应的超几何分布的乘积(Lancaster, 1949),给出关于公式 3.14 的分割。

3.37 说明为什么在公式 3.17 中 $\{n_{+j}\}$ 是 $\{\pi_{+j}\}$ 的充分统计量。

3.38 假设满足独立性,并且令 $p_{ij}=n_{ij}/n$ 以及 $\hat{\pi}_{ij}=p_{i+}p_{+j}$,

a. 证明对于 $\pi_{ij}=\pi_{i+}\pi_{+j}$,$p_{ij}$ 和 $\hat{\pi}_{ij}$ 是无偏的。

b. 证明 $\text{var}(p_{ij}) = \pi_{i+}\pi_{+j}(1-\pi_{i+}\pi_{+j})/n$。

c. 利用 $E(p_{i+}p_{+j})^2 = E(p_{i+}^2)(p_{+j}^2)$ 以及 $E(p_{i+}^2) = \text{var}(p_{i+}) + [E(p_{i+})]^2$,证明

$$\text{var}(\hat{\pi}_{ij}) = \{\pi_{i+}\pi_{+j}[\pi_{i+}(1-\pi_{+j}) + \pi_{+j}(1-\pi_{i+})]\}/n + \pi_{i+}(1-\pi_{i+})\pi_{+j}(1-\pi_{+j})/n^2。$$

d. 随着 $n\to\infty$,证明 $\lim \text{var}(\sqrt{n}\hat{\pi}_{ij}) \leqslant \lim \text{var}(\sqrt{n}p_{ij})$,并且只有当 $\pi_{ij}=1$ 或 0 时等式才成立。因而,如果模型成立或接近成立,模型估计值要优于样本比例。

3.39 证明不确定性系数(式 2.13)的样本值满足 $\hat{U} = -G^2/2n(\sum p_{+j}\log p_{+j})$(Haberman(1982)给出了它的标准误)。

3.40 当一个检验统计量具有连续分布时,P 值服从一个均匀零分布,对于 $0<\alpha<1$,$P(P\text{ 值}\leqslant\alpha)=\alpha$。针对 Fisher 精确检验,说明为什么在零假设下,对于 $0<\alpha<1$,$P(P\text{ 值}\leqslant\alpha)\leqslant\alpha$(提示:$P(P\text{ 值}\leqslant\alpha) = E[P(P\text{ 值}\leqslant\alpha \mid n_{1+}, n_{+1}, n)]$)。

3.41 参见注解 3.3 关于超几何分布(式 3.16)的矩量的介绍。令 $\rho = n_{+1}/n$,证明 n_{11} 具有与一个成功概率为 ρ 的 n_{1+} 次试验的二项随机变量相同的均值,并且它的方差乘上了一个有限总体校正因子 $(n-n_{1+})/(n-1)$(当 n_{1+} 比 n 小很多时,超几何分布与二项分布相似)。

3.42 由两个独立的二项分布变量形成的一个列联表分行的计数为(3,0/0,3)。对于 $H_0: \pi_1=\pi_2$ 和 $H_a: \pi_1>\pi_2$,证明精确无条件检验的 P 值等于 $\frac{1}{64}$,Fisher 精确检验的 P 值等于 $\frac{1}{20}$(有关的讨论,参见:Little(1989),Yates(1984)(G. Barnard 在后面的评论),Sprott(2000,Sec. 6.4.4))。

3.43 参见习题 3.42 以及通过 X^2 进行的关于 $H_a: \pi_1\neq\pi_2$ 的精确检验,说明为什么在 $\pi=0.5$ 时,无条件检验的 P 值与不同表格下的 Fisher 条件检验的 P 值之间的关系为

$$P(X^2\geqslant 6) = \sum_{k=0}^{6} P(X^2\geqslant 6 \mid n_{+1}=k)P(n_{+1}=k)。$$

因此,无条件 P 值$\left(\frac{1}{32}\right)$是一个对所观测到的列边际的 Fisher P 值以及对应着当其他的边际发生时不可能出现所观测到的极端结果的 P 值(0)的加权平均$\left(\text{即},\frac{1}{32}=0.10\left[\binom{6}{3}(1/2)^6\right]\right)$。第 3.5.6 节中所给出的 Fisher 的评论表明了他对此的看法。

3.44 考虑在给定边际的情况下关于 $I\times J$ 表格的独立性的精确检验,表格中对于 $i=1,\cdots,I$ 存在 $n_{ii}=1$,以及 $n_{ij}=0$。证明:(a)通过按照它们发生的概率对表格进行排序的检验,X^2 或 G^2 的 P 值等于 1.0;(b)按照一个定序统计量比如 r 或 $C-D$ 来对表格排序的单边检验的 P 值等于 $(1/I!)$。

3.45 一个蒙特卡洛程序随机抽取 M 个满足观测到的边际的不同 $I\times J$ 表格来近似精确检验中的 $P_o = P(X^2\geqslant X_o^2)$,令 $\hat{P}$ 表示这 M 个表格中满足 $X^2\geqslant X_o^2$ 的样本比例。证明 $P(|\hat{P}-P_o|\leqslant B) = 1-\alpha$ 要求 $M\approx z_{\alpha/2}^2 P_o(1-P_o)/B^2$。

3.46 对于公式 3.18 的分布,证明关于 θ 的条件最大似然估计满足 $n_{11}=E(n_{11})$。

4 广义线性模型简介

在第 2 章、第 3 章我们集中介绍了关于二维列联表的分析方法。然而,大多数研究会包括多个解释变量,并且这些变量既可能是连续的,也可能是分类的。研究的目的通常是描述这些解释变量对结果变量的效应。对这些效应*构建模型*(*Modeling*)能帮助我们有效地实现这一目标。可以通过对数据拟合很好的模型评估这些效应,包括相应的交互效应,并提供有关结果变量发生概率的修匀估计。

本书余下的部分将集中讨论关于分类结果变量的模型构建。在本章中,我们介绍一族*广义线性模型*(*generalized linear models*),它包括关于分类结果变量的最重要的模型以及关于连续结果变量的标准模型,第 4. 1 节介绍所有广义线性模型共有的三个组成部分。第 4. 2 节给出关于二分结果变量的模型,其中最重要的模型是 logistic *回归*(*logistic regression*),它是关于二项分布参数的 Logit 转换的线性模型。在第 5—7 章,我们将对这一模型展开更详细的讨论。

在第 4. 3 节,我们介绍关于计数数据的广义线性模型。*泊松回归模型*(*Poisson regression model*)是针对泊松分布变量均值的对数的线性模型,也称为*对数线性模型*(*loglinear model*)。关于利用该模型来分析列联表计数的方法,我们将在第 8 章和第 9 章介绍。

第 4. 4—4. 8 节的内容技术性相对较强。想对有关方法做一个大致了解的读者可以跳过或略读这些内容。第 4. 4 节的内容是广义线性模型的似然方程以及有关模型参数的最大似然估计的渐近协方差矩阵。第 4. 5 节综述有关统计推断的方法。第 4. 6 节介绍求解似然方程的方法。在最后两节,我们对本章中的模型做进一步扩展,讨论*类似然函数*(*quasi-likelihood*)以及*广义可加模型*(*generalized additive models*)。

4.1 广义线性模型

广义线性模型(*generalized linear models*, GLMs)扩展了普通的回归模型,可以分析非正态分布的结果变量以及相应均值的函数。广义线性模型包括三个组成部分:*随机部分*(*random component*)设定结果变量 Y 及其概率分布;*系统部分*(*systematic component*)设定线性预测函数中所使用的解释变量;以及*连结函数*(*link function*),即使模型的系统部分等于 $E(Y)$的函数。Nelder 和 Wedderburn(1972)最早提出了广义线性模型的概念,但在此之前,其所包括的许多模型就已经广为人知。

4.1.1 广义线性模型的组成部分

广义线性模型的*随机部分*(*random component*)由结果变量 Y 构成,该变量值(y_1,…,

y_N)为服从某一自然指数族分布的独立观测值。自然指数族(natural exponential family)分布的概率密度函数具有以下形式:

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_iQ(\theta_i)]。\tag{4.1}$$

包括泊松分布和二项分布等重要的分布都是它的特例。对于 $i=1,\cdots,N$,参数 θ_i 的值可以随解释变量的变动而变动。$Q(\theta)$项被称为*自然参数*(*natural parameter*)。在第4.4节,我们将给出同时包括离散参数的更一般化的公式。不过,对于基本的离散数据模型,公式4.1就足够了。

广义线性模型的*系统部分*(*systematic component*)通过线性模型将向量($\eta_1,\cdots,\eta_N$)与解释变量相联系。令 x_{ij}表示研究对象 i 的第 j 个预测变量的值($j=1,2,\cdots,p$),则有:

$$\eta_i = \sum_j \beta_j x_{ij}, \quad i = 1,\cdots,N。$$

这个解释变量的线性组合被称为*线性预测项*(*linear predictor*)。通常来说,在 x_{ij}中有一个对所有 i 的取值都等于1,它对应于模型中的截距(一般由 α 表示)。

广义线性模型的第三个组成部分是将随机部分和系统部分联系起来的*连结函数*(*link function*)。令 $\mu_i=E(Y_i)$, $i=1,\cdots,N$。模型通过 $\eta_i=g(\mu_i)$将 μ_i 和 η_i 相连结,其中连结函数 g 为单调可导函数。因此,通过以下公式,g 将 $E(Y_i)$与解释变量联系起来:

$$g(\mu_i) = \sum_j \beta_j x_{ij}, \quad i = 1,\cdots,N。\tag{4.2}$$

连结函数 $g(\mu) = \mu$ 被称为*恒等连结*(*identity link*)。这时,存在 $\eta_i = \mu_i$,即关于均值本身的线性模型。恒等连结就是Y服从正态分布时的标准回归模型的连结函数。*典型连结*(*canonical link*)是将均值转换为自然参数的连结函数。此时,$g(\mu_i) = Q(\theta_i)$,$Q(\theta_i) = \sum_j \beta_j x_{ij}$。相应的实例,见下一小节的介绍。

总的来说,广义线性模型是对服从自然指数分布的结果变量均值的函数的线性模型。接下来,我们将介绍有关离散结果变量的广义线性模型,并详释它的三个组成部分。

4.1.2 二分数据的二项 Logit 模型

常见的许多结果变量都是二分变量,取值1和0分别代表成功和失败。这种*伯努利试验*(*Bernoulli trial*)数据对应着*伯努利分布*(*Bernoulli distribution*),该分布设定了概率 $P(Y=1)=\pi$ 和 $P(Y=0)=1-\pi$,并有 $E(Y)=\pi$。这是二项分布(式1.1)在 $n=1$ 时的特例。它的概率密度函数为

$$\begin{aligned} f(y;\pi) &= \pi^y(1-\pi)^{1-y} = (1-\pi)[\pi/(1-\pi)]^y \\ &= (1-\pi)\exp\left(y\log\frac{\pi}{1-\pi}\right), \end{aligned}\tag{4.3}$$

其中 $y=0$ 或1。伯努利分布属于公式4.1中的自然指数族,其中 θ 等于 π,$a(\pi)=1-\pi$,$b(y)=1$,以及 $Q(\pi)=\log[\pi/(1-\pi)]$。自然参数 $\log[\pi/(1-\pi)]$是结果变量为1时的对数发生比,即 π 的 Logit。Logit 连结是一个典型连结。使用 Logit 连结的广义线性模型通常被称为 Logit *模型*(*Logit models*)。

4.1.3 计数数据的泊松对数线性模型

有些结果变量的可能取值是以计数的形式出现的。例如,对于生产电脑芯片所使用的硅片的样本,观测值可能是硅片上瑕疵的数量。另外,列联表中的数值也是计数。

关于计数数据的最简单分布是泊松分布。与计数相同,泊松变量可以取任意非负的

整数值。令 Y 表示某一计数变量，并令 $\mu=E(Y)$。关于 Y 的泊松概率密度函数(式 1.4)为

$$f(y;\mu)=\frac{e^{-\mu}\mu^{y}}{y!}=\exp(-\mu)\left(\frac{1}{y!}\right)\exp(y\log\mu),\quad y=0,1,2,\cdots。$$

它满足自然指数形式(式 4.1)，其中 $\theta=\mu, a(\mu)=\exp(-\mu), b(y)=1/y!$，以及 $Q(\mu)=\log\mu$。该分布的自然参数为 $\log\mu$，因而其典型连结函数是对数连结，即 $\eta=\log\mu$。使用对数连结的模型为

$$\log\mu_i=\sum_j\beta_j x_{ij},\quad i=1,\cdots,N。\tag{4.4}$$

该模型被称为*泊松对数线性模型*(*Poisson loglinear model*)。

4.1.4 连续结果变量的广义线性模型

广义线性模型也包括针对连续结果变量的模型。正态分布属于具有离散参数的自然指数族，它的自然参数为均值。因此，关于 $E(Y)$ 的普通回归模型就是一个使用恒等连结的广义线性模型。表 4.1 列出了这一模型以及其他包括正态随机部分的标准模型。此外，该表还列出了下面六章将介绍的有关离散结果变量的广义线性模型。

传统数据分析的做法是将 Y 进行转换，使其近似于同方差的正态分布，这样就可以使用普通最小二乘法回归。与之相反，在广义线性模型中，对连结函数的选取和对随机部分的设定是分开的。如果连结函数的意义在于使预测变量在模型中呈线性形式，那么它并不必然保证方差恒定或满足正态分布。这是因为拟合过程会最大化与 Y 的分布相对应的似然函数，而该分布并不局限于正态分布。

表 4.1 统计分析中广义线性模型的类型

随机部分	连结函数	系统部分	模型	章节
正态	恒等连结	连续	回归	
正态	恒等连结	分类	方差分析	
正态	恒等连结	混合	协方差分析	
二项	Logit 连结	混合	Logistic 回归	第 5,6 章
泊松	对数连结	混合	对数线性	第 8,9 章
多项	广义 Logit 连结	混合	多项结果	第 7 章

4.1.5 偏离度

就关于观测值 $\mathbf{y}=(y_1,\cdots,y_N)$ 的某一特定广义线性模型而言，令 $L(\boldsymbol{\mu};\mathbf{y})$ 为相应的对数似然函数，表示为均值 $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_N)$ 函数。令 $L(\hat{\boldsymbol{\mu}};\mathbf{y})$ 表示该模型的最大对数似然值。在所有可能的模型中，最大可能出现的对数似然值为 $L(\mathbf{y};\mathbf{y})$。这种情况出现在最一般性的模型中，即当每个观测值都对应着一个单独的参数时，存在完全拟合 $\hat{\boldsymbol{\mu}}=y$。这样的模型被称为*饱和模型*(*saturated model*)。该模型并不具有实际价值，因为它不对数据做任何简化。然而，饱和模型可以作为评价其他模型拟合情况的基准。

泊松或二项分布广义线性模型的*偏离度*(*deviance*)被定义为

$$-2[L(\hat{\boldsymbol{\mu}};\mathbf{y})-L(\mathbf{y};\mathbf{y})]。$$

这是一个似然比检验统计量，检验的零假设为，与一般情形(即，饱和模型)相比，该特定

模型是否成立。在一些泊松和二项分布广义线性模型中，当观测值的数量增加时，观测值取值组合的类别数 N 是不变的。这时，偏离度渐近服从卡方零分布。其自由度为 $df = N - p$，其中 p 是该模型所包括的参数个数，也即，自由度等于饱和模型与非饱和模型的参数数量之差。这种情况下，偏离度可以用作检验模型的拟合程度。

举例来说，在二项分布数据中，当试验的次数上升时，预测项取值组合的数量 N 保持不变。令 Y_i 服从 $\text{bin}(n_i, \pi_i)$ 分布，其中 $i = 1, \cdots, N$。考虑简单的同质性模型，即对所有 i，$\pi_i = \alpha$。该模型包括 $p = 1$ 个参数。饱和模型不对 $\{\pi_i\}$ 做任何假定，$\{\pi_i\}$ 可以取从 0 到 1.0间的任意 N 个值，也即包括 N 个参数。同质性模型的偏离度对应的自由度为 $df = N - 1$。事实上，该偏离度等于对这些样本所形成的 $N \times 2$ 表格进行独立性检验的似然比统计量 G^2（式 3.11）。在满足独立性的情况下，对于给定的 N，随着 $\{n_i\}$ 的增加，该统计量近似服从卡方分布。

在本书中，我们利用偏离度来检查模型的拟合情况，并对不同的模型加以比较。偏离度是由拟合不充分所对应的残差组成的。分析偏离度的方法是对有关正态线性模型的方差分析的扩展。

4.1.6 广义线性模型的优点

广义线性模型提供了一个关于模型分析的统一理论，它涵盖了有关连续变量和离散变量的重要模型。本书所介绍的模型是具有二项分布或泊松分布随机部分的广义线性模型，以及相应的多元扩展。这些模型可以利用第 4.6 节所介绍的算法进行最大似然参数估计，该算法通过对加权最小二乘法的结果进行迭代来求解。将广义线性模型限定于 Y 服从指数族分布的原因在于，无论选择哪一种连结函数，该算法适用于整个分布族。

大多数统计软件都可以拟合广义线性模型。有关细节，参见附录 A。

4.2 二分数据的广义线性模型

令 Y 表示一个二分的结果变量。例如，Y 可以是某次英国大选中的投票结果（工党，保守党），对汽车的选择（国产，进口），或者乳腺癌的诊断结果（有，没有）。每个观测值都是两种可能结果中的一种，由 0 或 1 来表示，每一次试验都服从二项分布。它的均值为 $E(Y) = P(Y = 1)$。我们用 $\pi(\mathbf{x})$ 来表示 $P(Y = 1)$，以反映它取决于预测变量 $\mathbf{x} = (x_1, \cdots, x_p)$ 的值。Y 的方差等于

$$\text{var}(Y) = \pi(\mathbf{x})[1 - \pi(\mathbf{x})],$$

即试验次数为一次的二项分布方差。在介绍有关二分数据的广义线性模型时，为了简化起见，我们考虑只存在一个解释变量的情形。

4.2.1 线性概率模型

对于二分结果变量，回归模型

$$\pi(x) = \alpha + \beta x \tag{4.5}$$

被称作线性概率模型（*linear probability model*）。当观测值之间相互独立时，该模型是由二项分布随机部分和恒等连结函数组成的广义线性模型。

线性概率模型存在一个重大的结构性缺陷。概率只能取 0 到 1 之间的值，而线性函数却可以取任意的实数值。当 x 值很大或很小时，模型（式 4.5）会出现 $\pi(x) < 0$ 或 $\pi(x) > 1$的情况。在模型包括多个预测变量的情况下，拟合相应模型常常会出现问题，这

是因为在拟合过程中对于某些 **x** 值,其所对应的 $\hat{\pi}(\mathbf{x})$ 会落在[0,1]区间之外。因此,该模型仅对某个限定范围的 x 值是有效的。当它成立时,其优点在于解释起来非常简单:β 表示 x 每增加一个单位所对应的 $\pi(x)$ 的变动。

关于拟合这个模型以及其他广义线性模型的技术细节,将留在第 4.6 节详细介绍。在这儿,我们仅强调应当假定 Y 服从二项分布,并应用最大似然估计而不是普通最小二乘法。在满足同方差正态分布的情况下,最小二乘法估计等价于最大似然估计。对于二分结果变量,使得最小二乘法估计最优(即,在一系列线性无偏估计值中方差最小)的同方差条件是无法满足的。由于 $\text{var}(Y)=\pi(x)[1-\pi(x)]$,方差通过 x 对 $\pi(x)$ 的影响而依赖于 x 的取值。当 $\pi(x)$ 的值趋近于 0 或 1 时,Y 的分布集中在一个点上,且它的方差趋近于 0。由于存在变动的方差,二项分布的最大似然估计比最小二乘法估计更有效。同时,由于 Y 是二分的,远远不同于正态分布,因此,最小二乘法估计常用的抽样分布在这里并不适用。然而,当样本 x 值所对应的 $\hat{\pi}(x)$ 的值落在一个方差相对稳定的区间(大约 0.3 到 0.7)时,最大似然估计和最小二乘法估计通常具有相似的估计值和标准误。

4.2.2 例子:打鼾与心脏病

我们通过表 4.2 来展示线性概率模型,该表取自一项对 2 484 个研究对象的流行病调查数据,以考察打鼾是否会导致心脏病。该表按照配偶关于被调查对象打鼾情况的报告进行了划分。模型设定,得心脏病的概率与打鼾的频繁程度 x 线性相关。我们将表中的各行视为独立的二项分布样本。对于 x 的类别,不存在明显的可选赋值。我们使用(0,2,4,5)作为 x 的值,以反映最后两组之间的距离比其他相邻组距离要小(习题 4.4 使用了等距的赋值)。无论我们使用原始数据文件(2 484 个二项分布观测值)还是使用表 4.2 中分别回答"是"与"否"的四个二项分布计数,所得到的最大似然估计值以及标准误都一样。

表 4.2 打鼾与心脏病的关系

打鼾	心脏病		回答"是"的比例	线性拟合结果[a]	Logit 拟合结果[a]
	是	否			
从不	24	1 355	0.017	0.017	0.021
偶尔	35	603	0.055	0.057	0.044
几乎每晚	21	192	0.099	0.096	0.093
每晚	30	224	0.118	0.116	0.132

a 模型拟合结果对应于回答"是"的比例。

来源:P. G. Norton and E. V. Dunn, *British Med. J.* 291:630-632(1985), BMJ Publishing Group.

软件(如 SAS,参见表 A.3)所给出的最大似然拟合为 $\hat{\pi}(x)=0.017\,2+0.019\,8x$,其中 $\hat{\beta}=0.019\,8$ 的标准误为 SE $=0.002\,8$。对于不打鼾的人($x=0$),模型所估计的得心脏病的比例为 0.017 2。我们将广义线性模型中 $E(Y)$ 的估计值称为*拟合值*(*fitted values*)。表 4.2 给出了样本比例和该模型的拟合值。相应地,图 4.1 对样本值和拟合值进行了绘图。这些结果都显示模型拟合得很好(对二分结果变量广义线性模型的拟合优度的正式分析,我们将留在第 5.2.3 节讨论)。对模型的解释很简单。对于非打鼾者,所估计的得心脏病的概率约为 0.02;对于偶尔打鼾者,估计的概率上升了 $2(0.019\,8)=0.04$;对于几乎每晚都打鼾的人,其概率又上升 0.04;对于一直打鼾者,其概率进一步上升 0.02。

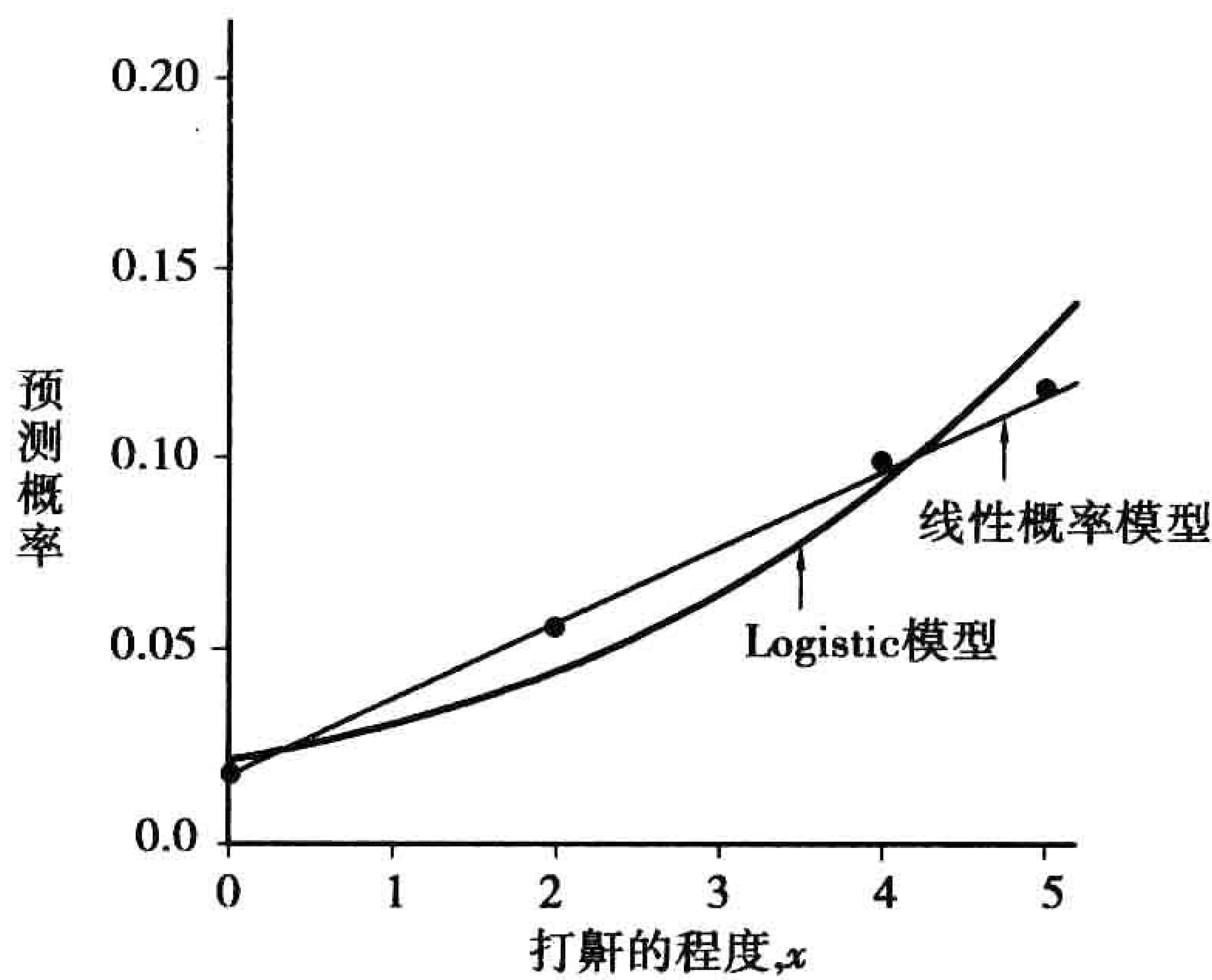


图 4.1 线性概率模型和 logistic 回归模型的预测概率

4.2.3 Logistic 回归模型

通常,二分数据源于 $\pi(x)$ 和 x 之间的*非线性*(*nonlinear*)关系。当 $\pi(x)$ 接近 0 或 1 时,x 变动相同单位所产生的影响往往比当 $\pi(x)$ 等于 0.5 时要小。以购买小汽车时考虑选择买新车或者二手车的情况为例。令 $\pi(x)$ 表示当家庭年收入 $=x$ 时选择购买新车的概率。家庭年收入同样增加 50 000 美元,在 $x=1\ 000\ 000$ 美元时(这时的 $\pi(x)$ 接近于 1)所产生的影响明显比在 $x=50\ 000$ 美元时小。

在实际应用中,$\pi(x)$ 和 x 之间的非线性关系往往是单调的,也即,随着 x 的增加,$\pi(x)$ 持续上升或者持续下降。图 4.2 中的 S 形曲线是一种很典型的情形。具有这一形状的最重要曲线对应的模型公式为

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}。 \tag{4.6}$$

这就是 logistic *回归*(*logistic regression*)模型。随着 $x \to \infty$,当 $\beta < 0$ 时 $\pi(x) \downarrow 0$,当 $\beta > 0$ 时 $\pi(x) \uparrow 1$。

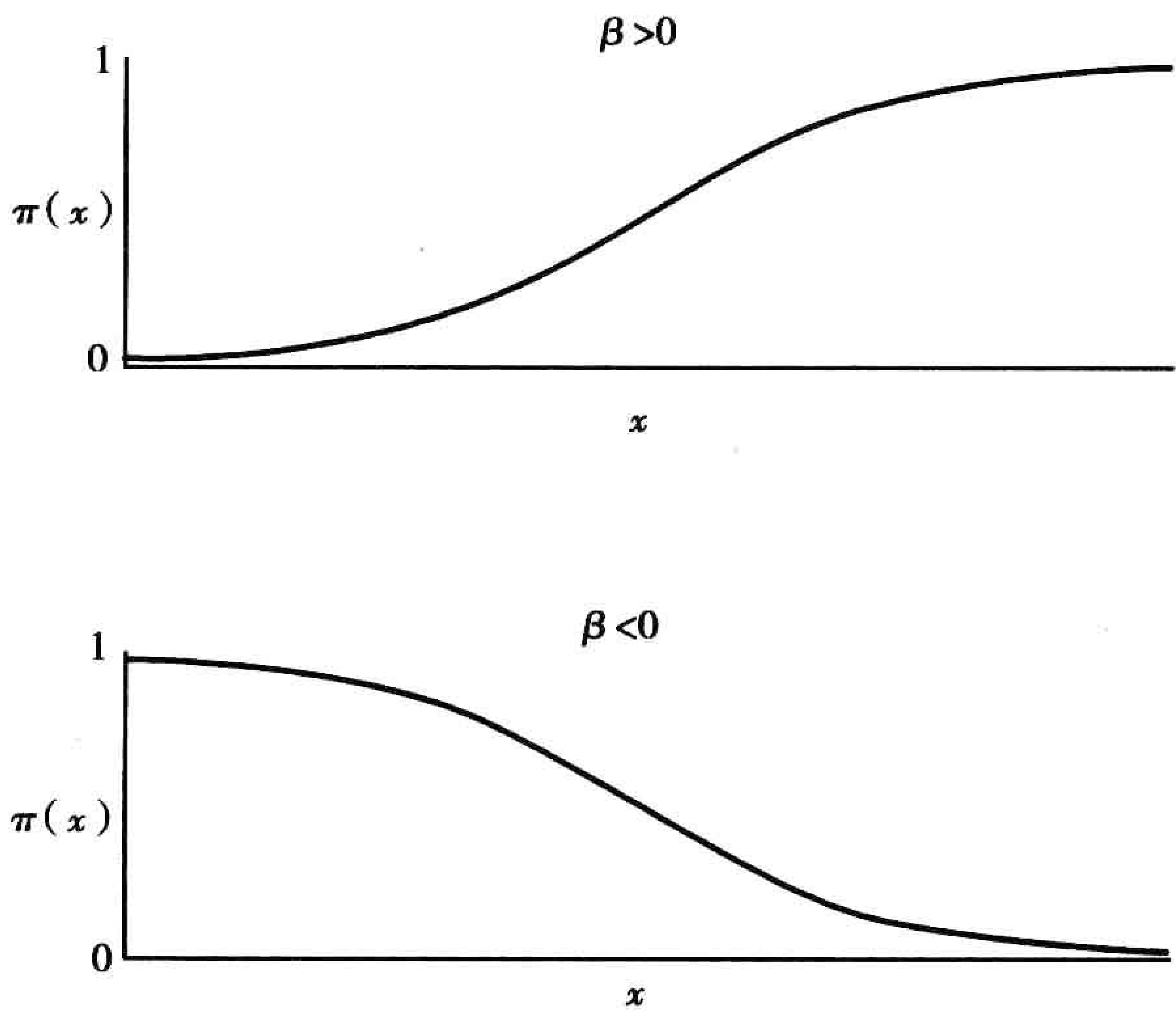


图 4.2 Logistic 回归函数

现在,让我们找出 logistic 回归作为一个广义线性模型的连结函数。对于公式 4.6,

发生比为

$$\frac{\pi(x)}{1-\pi(x)}=\exp(\alpha+\beta x)。$$

对数发生比具有线性关系

$$\log\frac{\pi(x)}{1-\pi(x)}=\alpha+\beta x。\tag{4.7}$$

因此,相应的连结函数为对数发生比转换,即 Logit。Logistic 回归模型是由二项随机部分和 Logit 连结函数组成的广义线性模型。因而,Logistic 回归模型也称为 Logit *模型*(*Logit models*)。

Logit 是二项分布的自然参数,所以 Logit 连结是它的典型连结。尽管 $\pi(x)$ 的值必须落入(0,1)区间,但 Logit 却可以取任意实数值。这也对应于构成广义线性模型系统部分的线性预测项(如 $\alpha+\beta x$)的取值范围。因此,这个模型不存在线性概率模型所固有的结构性问题。

用表 4.2 所示的打鼾数据,统计软件给出的 Logistic 回归的最大似然拟合为

$$\text{Logit}[\hat{\pi}(x)]=-3.87+0.40x。$$

$\hat{\beta}=0.40$ 为正值表明,较频繁的打鼾伴随着心脏病发病率的上升。在第 5 章和第 6 章,我们将具体地讨论 logistic 回归,并对上述公式进行解释。将 x 的值代入概率估计公式(式 4.6),可以得到估计的概率值。表 4.2 给出了该模型的拟合值,图 4.1 则显示了相应的拟合情况。在这里,估计概率的取值范围很小,所以它的拟合接近于线性,而且其结果与线性概率模型的结果很相似。

4.2.4 关于 2×2 列联表的二项分布广义线性模型

在二分结果变量的广义线性模型中,最简单的情况是模型只包括一个解释变量 X,并且 X 也是二分变量,将它的值分别记为 0 和 1。对于给定的连结函数,广义线性模型

$$连结[\pi(x)]=\alpha+\beta x$$

中 X 的效应可以表示为

$$\beta=连结[\pi(1)]-连结[\pi(0)]。$$

在恒等连结的情况下,$\beta=\pi(1)-\pi(0)$,X 的效应等于比例之差。在对数连结的情况下,$\beta=\log[\pi(1)]-\log[\pi(0)]=\log[\pi(1)/\pi(0)]$,$X$ 的效应等于相对风险的对数。对于 Logit 连结,

$$\beta=\text{Logit}[\pi(1)]-\text{Logit}[\pi(0)]=\log\frac{\pi(1)}{1-\pi(1)}-\log\frac{(0)}{1-\pi(0)}$$
$$=\log\frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))},$$

X 的效应等于对数发生比之比。在 2×2 表格中,关联度量指标就是有关二分数据的广义线性模型的效应参数。

4.2.5 Probit 与累积分布函数(CDF)的反函数连结*

单调回归曲线与连续随机变量的累积分布函数(cdf)的形状相一致,如图 4.2 中的第一条曲线。这表明,对于某一累积分布函数 F,二分结果变量模型具有 $\pi(x)=F(x)$ 的形式。

利用一组完整的的累积分布函数,如具有不同均值和方差的正态累积分布函数,可

以实现曲线 $\pi(x)=F(x)$ 在增长率以及最快增长位置方面的灵活性。令 $\Phi(\cdot)$ 表示一组分布的标准累积分布函数,如 $N(0,1)$ 的累积分布函数。通过 Φ,模型可表示为

$$\pi(x) = \Phi(\alpha+\beta x), \tag{4.8}$$

它提供了同样的灵活性。随着 α 和 β 的变化,会出现不同累积分布函数的形状。用 βx 替代 x 允许曲线在标准累积分布函数的基础上以不同的速度上升(或者下降,对应于 $\beta<0$ 时);α 的不同取值允许曲线向左或向右平移。

当 Φ 在整个实数范围内严格上升时,它存在反函数 $\Phi^{-1}(\cdot)$,并且公式 4.8 等价于

$$\Phi^{-1}[\pi(x)] = \alpha+\beta x。\tag{4.9}$$

对于这组累积分布函数,广义线性模型的连结函数为 Φ^{-1}。连结函数将概率的取值范围 $(0,1)$ 对应扩展到 $(-\infty,\infty)$,即线性预测项的取值范围。特别地,当 Φ 是标准正态分布的累积分布函数时,曲线具有正态累积分布函数的形状。这时,模型(式 4.9)被称作 probit 模型。这个曲线的形状和 logistic 回归曲线相似。我们将在第 6.6 节讨论 probit 模型。

当 $\beta>0$ 时,式 4.6 的 logistic 回归曲线是 logistic 分布(*logistic distribution*)的累积分布函数。当 $\beta<0$ 时,$1-\pi(x)$ 或概率 $Y=0$ 的曲线具有以上的形状。均值为 μ、离散参数为 $\tau>0$ 的 logistic 分布的累积分布函数为

$$F(x) = \frac{\exp[(x-\mu)/\tau]}{1+\exp[(x-\mu)/\tau]}, \quad -\infty < x < \infty。$$

它所对应的概率密度函数呈对称的钟形,其标准差等于 $\tau\pi/\sqrt{3}$(这里,π 为数学常数 3.14…)。它的形状与具有相同均值和标准差的正态密度函数非常像,只是尾部稍微高一些(它的峰度与自由度 df=9 的 t 分布相同)。

标准化形式的 logistic 累积分布函数具有 $\mu=0$ 以及 $\tau=1$,所以 $\Phi(x)=e^x/(1+e^x)$。对于该函数,logistic 回归曲线(式 4.6)的形式为 $\pi(x)=\Phi(\alpha+\beta x)$。按照(式 4.9),Logit 转换实际上是标准 logistic 累积分布函数的反函数,也即,当 $\Phi(x)=\pi(x)=e^x/(1+e^x)$ 时,$x=\Phi^{-1}[\pi(x)]=\log[\pi(x)/(1-\pi(x))]$。

4.3 计数数据的广义线性模型

最常用的有关计数数据的广义线性模型假定 Y 服从泊松分布。我们在第 1.2.3 节介绍了这个分布。在第 8 章和第 9 章,我们将讨论由分类结果变量组成的列联表计数的泊松广义线性模型。在本节中,我们介绍泊松广义线性模型的另一种应用,即对一个单一离散结果变量的计数或比率数据构建模型。

4.3.1 泊松对数线性模型

泊松分布具有正的均值 μ。尽管广义线性模型可以通过恒等连结来对正的均值建立模型,但是更通行的做法是对均值的对数构建模型。与线性预测项 $\alpha+\beta x$ 一致,均值的对数可以取任意实数值。均值的对数是泊松分布的自然参数,因而对数连结是泊松广义线性模型的典型连结。泊松对数线性模型使用对数连结,并假定 Y 服从泊松分布。

以 X 为解释变量的泊松对数线性模型可表示为

$$\log\mu = \alpha+\beta x。\tag{4.10}$$

在该模型中,均值满足指数关系

$$\mu = \exp(\alpha+\beta x) = e^{\alpha}(e^{\beta})^{x}。\tag{4.11}$$

x 每增加一个单位,对 μ 产生的影响具有 e^{β} 的乘数效应,即在 $x+1$ 处的均值等于 x 处的

均值再乘以 e^{β}。

4.3.2 例子:马蹄蟹的同伴

现在,我们通过表 4. 3 中关于巢居马蹄蟹的一项研究来具体说明泊松广义线性模型。每一只雌性马蹄蟹在它的巢穴内都有一只雄性马蹄蟹。该研究考察在雌性马蹄蟹的巢穴附近,是否还有其他雄性马蹄蟹(称为同伴)居住的影响因素。解释变量包括雌性马蹄蟹的颜色、蟹刺状况、体重以及壳宽,结果变量为每只雌性马蹄蟹的同伴数量。就当前的例子,我们使用壳宽作为唯一的预测变量。表 4. 3 列出了以厘米为单位的壳宽测量值,其样本均值为 26. 3,标准差为 2. 1。

表 4. 3 按照雌性马蹄蟹特征划分的同伴数[a]

C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa
2	3	28.3	3.05	8	3	3	22.5	1.55	0	1	1	26.0	2.30	9	3	3	24.8	2.10	0
3	3	26.0	2.60	4	2	3	23.8	2.10	0	3	2	24.7	1.90	0	2	1	23.7	1.95	0
3	3	25.6	2.15	0	3	3	24.3	2.15	0	2	3	25.8	2.65	0	2	3	28.2	3.05	11
4	2	21.0	1.85	0	2	1	26.0	2.30	14	1	1	27.1	2.95	8	2	3	25.2	2.00	1
2	3	29.0	3.00	1	4	3	24.7	2.20	0	2	3	27.4	2.70	5	2	2	23.2	1.95	4
1	2	25.0	2.30	3	2	1	22.5	1.60	1	3	3	26.7	2.60	2	4	3	25.8	2.00	3
4	3	26.2	1.30	0	2	3	28.7	3.15	3	2	1	26.8	2.70	5	4	3	27.5	2.60	0
2	3	24.9	2.10	0	1	1	29.3	3.20	4	1	3	25.8	2.60	0	2	2	25.7	2.00	0
2	1	25.7	2.00	8	2	1	26.7	2.70	5	4	3	23.7	1.85	0	2	3	26.8	2.65	0
2	3	27.5	3.15	6	4	3	23.4	1.90	0	2	3	27.9	2.80	6	3	3	27.5	3.10	3
1	1	26.1	2.80	5	1	1	27.7	2.50	6	2	1	30.0	3.30	5	3	1	28.5	3.25	9
3	3	28.9	2.80	4	2	3	28.2	2.60	6	2	3	25.0	2.10	4	2	3	28.5	3.00	3
2	1	30.3	3.60	3	4	3	24.7	2.10	5	2	3	27.7	2.90	5	1	1	27.4	2.70	6
2	3	22.9	1.60	4	2	1	25.7	2.00	5	2	3	28.3	3.00	15	2	3	27.2	2.70	3
3	3	26.2	2.30	3	2	1	27.8	2.75	0	4	3	25.5	2.25	0	3	3	27.1	2.55	0
3	3	24.5	2.05	5	3	1	27.0	2.45	3	2	3	26.0	2.15	5	2	3	28.0	2.80	1
2	3	30.0	3.05	8	2	3	29.0	3.20	10	2	3	26.2	2.40	0	2	1	26.5	1.30	0
2	3	26.2	2.40	3	3	3	25.6	2.80	7	3	3	23.0	1.65	1	3	3	23.0	1.80	0
2	3	25.4	2.25	6	3	3	24.2	1.90	0	2	2	22.9	1.60	0	3	2	26.0	2.20	3
2	3	25.4	2.25	4	3	3	25.7	1.20	0	2	3	25.1	2.10	5	3	2	24.5	2.25	0
4	3	27.5	2.90	0	3	3	23.1	1.65	0	3	1	25.9	2.55	4	2	3	25.8	2.30	0
4	3	27.0	2.25	3	2	3	28.5	3.05	0	4	1	25.5	2.75	0	4	3	23.5	1.90	0
2	2	24.0	1.70	0	2	1	29.7	3.85	5	2	1	26.8	2.55	0	4	3	26.7	2.45	0
2	1	28.7	3.20	0	3	3	23.1	1.55	0	2	1	29.0	2.80	1	3	3	25.5	2.25	0
3	3	26.5	1.97	1	3	3	24.5	2.20	1	3	3	28.5	3.00	1	2	3	28.2	2.87	1
2	3	24.5	1.60	1	2	3	27.5	2.56	1	2	2	24.7	2.55	4	2	1	25.2	2.00	1
3	3	27.3	2.90	1	2	3	26.3	2.40	1	2	3	29.0	3.10	1	2	3	25.3	1.90	2
2	3	26.5	2.30	4	2	3	27.8	3.25	3	2	3	27.0	2.50	6	3	3	25.7	2.10	0
2	3	25.0	2.10	2	2	3	31.9	3.33	2	4	3	23.7	1.80	0	4	3	29.3	3.23	12
3	3	22.0	1.40	0	2	3	25.0	2.40	5	3	3	27.0	2.50	6	3	3	23.8	1.80	6
1	1	30.2	3.28	2	3	3	26.2	2.22	0	2	3	24.2	1.65	2	2	3	27.4	2.90	3

续表

C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa
2	2	25.4	2.30	0	3	3	28.4	3.20	3	4	3	22.5	1.47	4	2	3	26.2	2.02	2
2	1	24.9	2.30	6	1	2	24.5	1.95	6	2	3	25.1	1.80	0	2	1	28.0	2.90	4
4	3	25.8	2.25	10	2	3	27.9	3.05	7	2	3	24.9	2.20	0	2	1	28.4	3.10	5
3	3	27.2	2.40	5	2	2	25.0	2.25	6	2	3	27.5	2.63	6	2	1	33.5	5.20	7
2	3	30.5	3.32	3	3	3	29.0	2.92	3	2	1	24.3	2.00	0	2	3	25.8	2.40	0
4	3	25.0	2.10	8	2	1	31.7	3.73	4	2	3	29.5	3.02	4	3	3	24.0	1.90	10
2	3	30.0	3.00	9	2	3	27.6	2.85	4	2	3	26.2	2.30	0	2	1	23.1	2.00	0
2	1	22.9	1.60	0	4	3	24.5	1.90	0	2	3	24.7	1.95	4	2	3	28.3	3.20	0
2	3	23.9	1.85	2	3	3	23.8	1.80	0	3	2	29.8	3.50	4	2	3	26.5	2.35	4
2	3	26.0	2.28	3	2	3	28.2	3.05	8	4	3	25.7	2.15	0	2	3	26.5	2.75	7
2	3	25.8	2.20	0	3	3	24.1	1.80	0	3	3	26.2	2.17	2	3	3	26.1	2.75	3
3	3	29.0	3.28	4	1	1	28.0	2.62	0	4	3	27.0	2.63	0	2	2	24.5	2.00	0
1	1	26.5	2.35	0															

a C,颜色(1,浅褐色;2,褐色;3,深褐色;4,黑色)。S,蟹刺状况(1,都很好;2,一个磨损或折断;3,两个都磨损或折断)。W,壳宽(厘米)。Wt,体重(千克)。Sa,同伴数。

来源:由佛罗里达大学动物学系的 Jane Brockmann 友情提供,该研究发表在:*Ethology* 102;1-21(1996)。

图 4.3 显示了结果变量(同伴)的计数和壳宽的关系,图中的数字符号表示落在每一点上的观测值的数量。该图中变量取值的变动性很大,因而很难清楚分辨两个变量之间的关系。为了更清楚地描述这一关系,我们按照雌性马蹄蟹的壳宽将其进行了分组($\leqslant$23.25,23.25 ~ 24.25,24.25 ~ 25.25,25.25 ~ 26.25,26.25 ~ 27.25,27.25 ~ 28.25,28.25 ~ 29.25,>29.25),并计算了每组雌性马蹄蟹的同伴数量的样本均值。图 4.4 显示了各组马蹄蟹的同伴数量样本均值与壳宽之间的关系。

描述趋势的更精确办法不是将壳宽的值分组或假定某种特定的函数关系,而是对数据进行修匀处理。图 4.4 也给出了一条基于对广义线性模型的扩展(见第 4.8 节)所得到的修匀曲线。样本均值和修匀曲线都显示出很强的上升趋势(由于每一组别中结果变量的计数倾向于向右偏斜,所以均值倾向于落在修匀曲线的上方;修匀曲线受奇异值的影响较小)。这个趋势看上去近似于线性,接下来我们讨论在未分组数据中均值或其对数与壳宽呈线性关系的模型。

对于雌性马蹄蟹,令 μ 为同伴数的期望值,x 表示壳宽。通过广义线性模型软件(如 SAS,参见附表 A.4),泊松对数线性模型(式 4.10)的最大似然估计为

$$\log \hat{\mu} = \hat{\alpha} + \hat{\beta}x = -3.305 + 0.164x。$$

壳宽的效应 $\hat{\beta}=0.164$ 是正的,其标准误 SE = 0.020。在任意的壳宽水平上,模型的拟合值为同伴数的估计均值 $\hat{\mu}$。例如,当壳宽取均值即 $x=26.3$ 时,相应的拟合值为

$$\hat{\mu} = \exp(\hat{\alpha} + \hat{\beta}x) = \exp[-3.305 + 0.164(26.3)] = 2.74。$$

对于这一模型,$\exp(\hat{\beta}) = \exp(0.164) = 1.18$ 是 x 每增加 1 厘米对 $\hat{\mu}$ 的乘数效应。例如,在 $x=27.3=26.3+1$ 时的拟合值为 $\exp[-3.305+0.164(27.3)]=3.23$,它等于 1.18×2.74,也即,壳宽每增加 1 厘米会导致估计的均值上升 18%。

图 4.4 显示 $E(Y)$ 可能随着壳宽呈线性上升趋势。这表明,可以考虑使用恒等连结

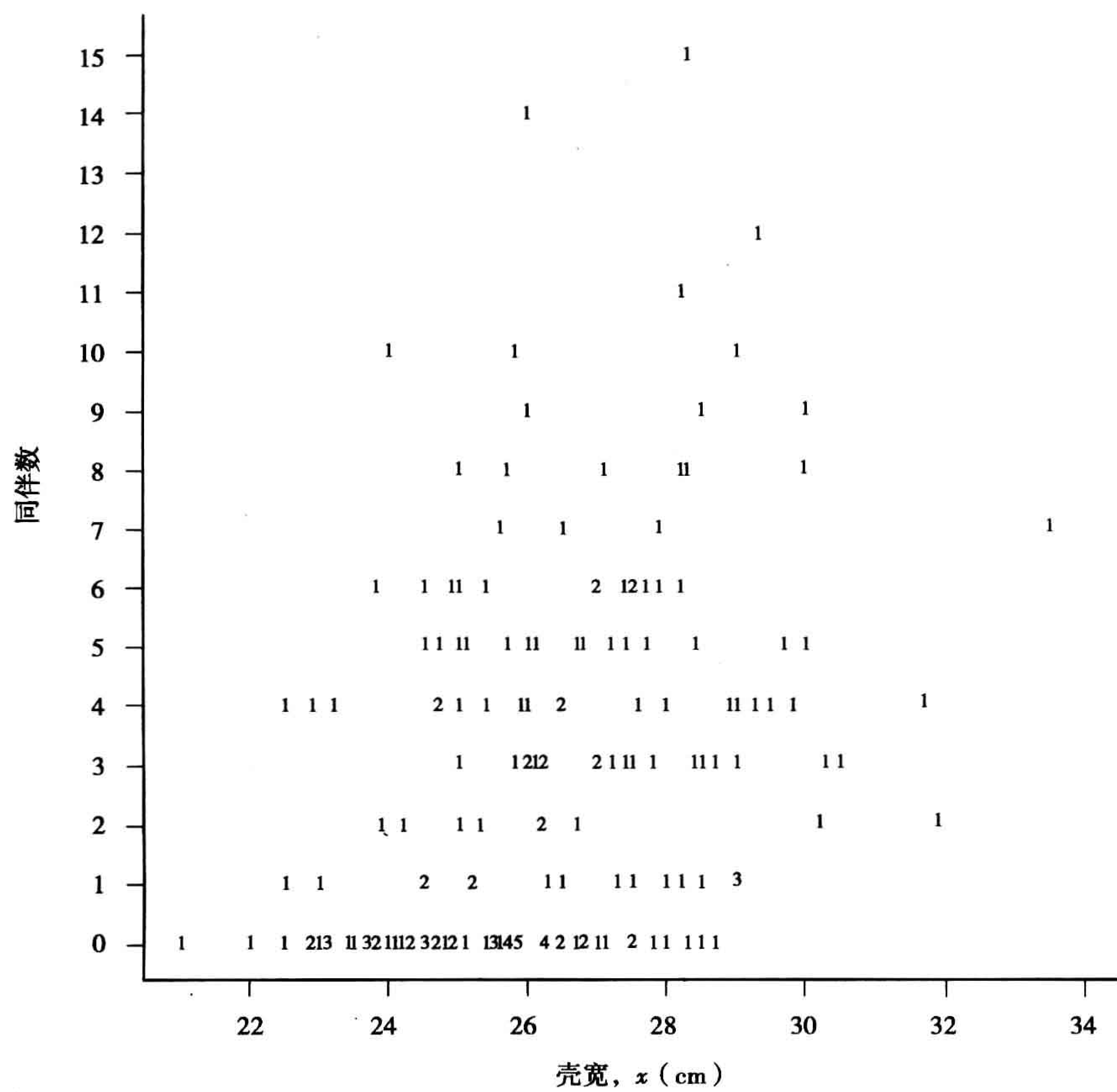


图 4.3 雌性马蹄蟹的壳宽与同伴数

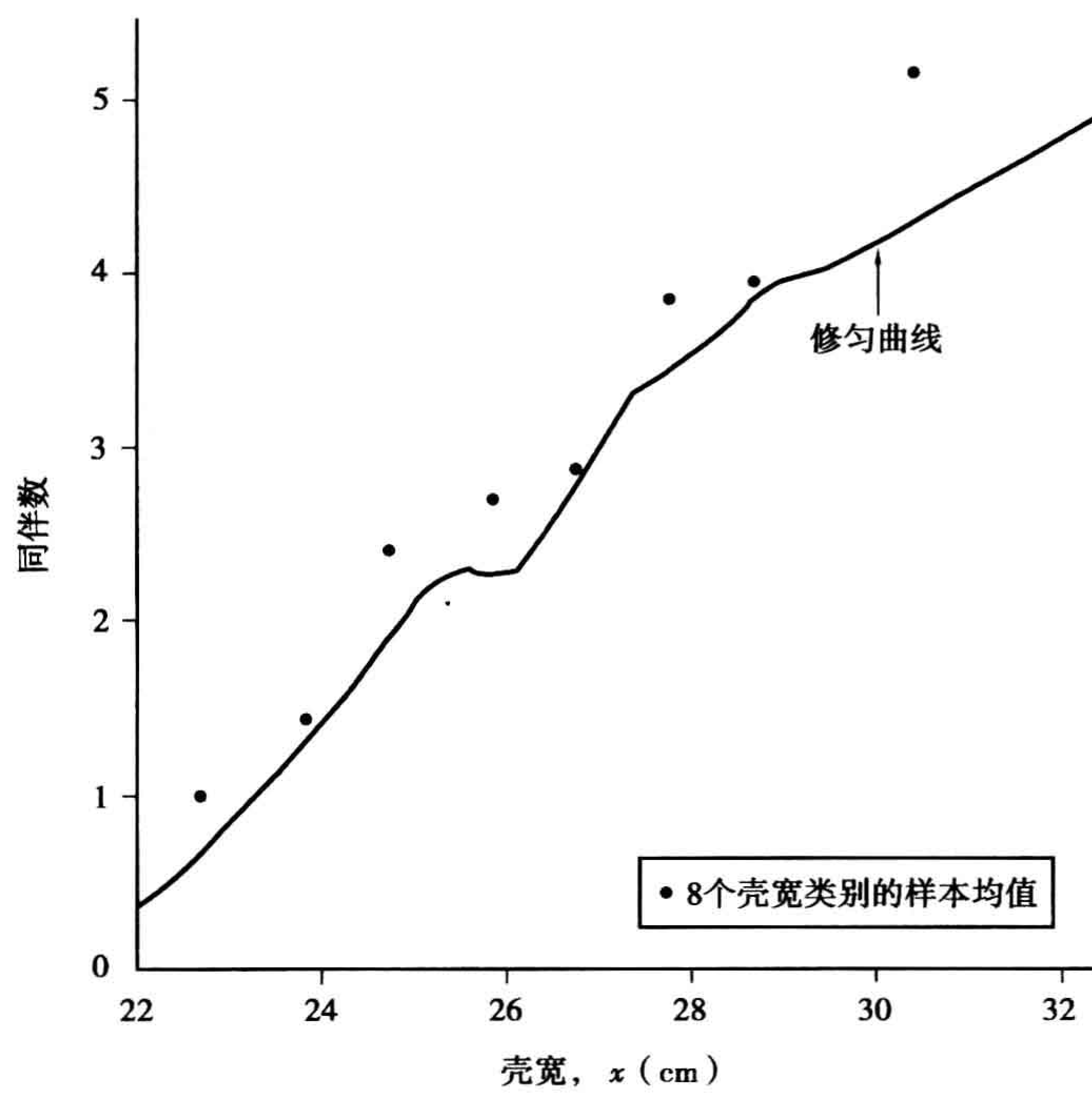


图 4.4 马蹄蟹计数的修匀曲线

的泊松广义线性模型。它的最大似然拟合为

$$\hat{\mu} = \hat{\alpha} + \hat{\beta}x = -11.53 + 0.55x。$$

这个模型中 X 对 μ 的影响是加数的而不是乘数的。x 每增加 1 厘米，会导致估计值 $\hat{\mu}$ 上升 $\hat{\beta}=0.55$。拟合值在所有 x 的样本取值范围内都为正，这个模型只是描述一种效应：平均来说，壳宽每增加大约 2 厘米，相应的同伴数就会增加一个。

图 4.5 分别显示了在对数连结和恒等连结模型中 $\hat{\mu}$ 与壳宽的关系。尽管当壳宽的值很小或很大时它们并不一致，但是在大多数观测值发生的壳宽范围内，二者的预测结果很相近。接下来，我们分析是否每个模型都对数据拟合得很充分。

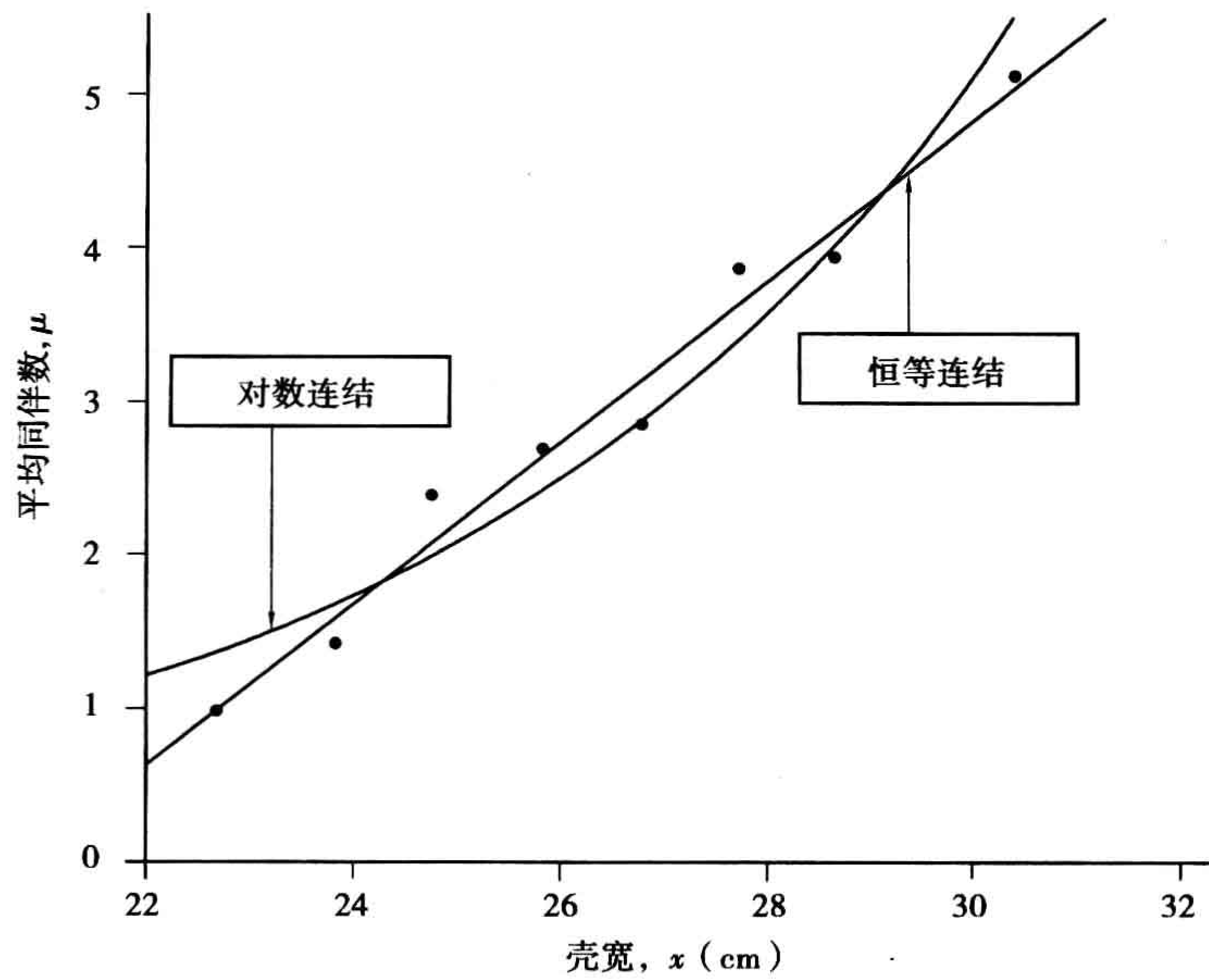


图 4.5 使用对数连结和恒等连结时有关同伴数的估计均值

4.3.3 泊松广义线性模型的过度离散

在第 1.2.4 节中我们提到，计数数据常常表现出比泊松分布所允许的更大的变动性。对于分组的马蹄蟹数据，表 4.4 给出了在每一壳宽类别内雌性马蹄蟹同伴数的样本均值和方差。这些方差比均值大得多，然而泊松分布对应的均值和方差相等。这种比广义线性模型的随机部分所预测的变动性更大的情形表明*存在过度离散*（*overdispersion*）。

表 4.4 同伴数的样本均值和方差

壳宽(cm)	马蹄蟹数量	同伴数	样本均值	样本方差
<23.25	14	14	1.00	2.77
23.25 ~ 24.25	14	20	1.43	8.88
24.25 ~ 25.25	28	67	2.39	6.54
25.25 ~ 26.25	39	105	2.69	11.38
26.25 ~ 27.25	22	63	2.86	6.88
27.25 ~ 28.25	24	93	3.87	8.81
28.25 ~ 29.25	18	71	3.94	16.88
>29.25	14	72	5.14	8.29

过度离散的一个常见原因是研究对象的异质性。例如，假设壳宽、体重、颜色和蟹刺状况是影响雌性马蹄蟹的同伴数的四个预测变量。假定对于预测变量的每一给定取值

组合,Y 服从泊松分布。我们前面的模型使用壳宽作为唯一的预测变量。这时,具有一定壳宽的马蹄蟹是由具有不同体重、颜色和蟹刺状况的马蹄蟹混合而成的。因此,具有给定壳宽的马蹄蟹的总体是多个泊松分布总体的混合体,其中每一个分布对结果变量而言都对应着自己特定的均值。这种异质性导致在给定壳宽的情况下,总的结果变量的分布比泊松分布所预测的变动性更大。假如在所有的相应变量都被控制后方差会等于均值,那么当仅仅控制一个变量时,方差会大于均值。

对于 Y 服从正态分布的普通回归,过度离散不是一个问题,因为正态分布有专门的参数(方差)来描述变动性。然而,对于二项分布和泊松分布,方差是均值的函数。在对计数数据进行模型分析时,过度离散是很常见的问题。当关于均值的模型是正确的,但真实的分布不服从泊松分布时,对模型参数的最大似然估计仍然是一致的,但是相应的标准误并不正确。接下来,我们介绍泊松广义线性模型的一种扩展,通过加入一个额外的参数从而较好地处理过度离散的问题。在第 4.7 节中,我们将介绍处理这一问题的另一种方法,即类似然(quasi-likelihood)统计推断。

4.3.4 负二项广义线性模型

负二项分布(*negative binomial distribution*)的概率密度函数为

$$f(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y, \quad y = 0,1,2\cdots, \qquad (4.12)$$

其中,k 和 μ 是参数。这个分布具有

$$E(Y) = \mu, \quad \text{var}(Y) = \mu + \mu^2/k。$$

指数 k^{-1} 被称为*离散参数*(*dispersion parameter*)。当 $k^{-1}\to 0$ 时,$\text{var}(Y)\to\mu$,相应的负二项分布收敛于泊松分布(Cameron and Trivedi, 1998, p. 75)。通常情况下,k^{-1} 是未知的,对该参数进行估计有助于我们了解过度离散的程度。

对于给定的 k,我们可以将式 4.12 表示为如式 4.1 的自然指数族的形式。这样,具有负二项分布随机部分的模型也属于广义线性模型。简单起见,在相应的模型中,尽管 k 是未知的,但对于所有的观测值设其为一个不变的常数。与二分数据的广义线性模型相类似,负二项广义线性模型也可以使用不同的连结函数。最常用的连结函数为对数连结,这一点与泊松对数线性模型相同。但是有时候,负二项广义线性模型也可以使用恒等连结。

我们将在第 13.4 节详细讨论负二项广义线性模型。在这里,我们利用前面在介绍泊松广义线性模型时使用过的马蹄蟹数据对其加以演示。使用恒等连结、以壳宽作为预测变量,拟合的泊松广义线性模型为 $\hat{\mu} = -11.53 + 0.55x$($\hat{\beta}$ 的标准误为 SE = 0.06),相应的负二项广义线性模型为 $\hat{\mu} = -11.15 + 0.53x$($\hat{\beta}$ 的标准误为 SE = 0.11)。其中,$\hat{k}^{-1}$ = 0.98。对于某一预测值 $\hat{\mu}$,泊松广义线性模型中的方差为 $\hat{\mu}$,而负二项广义线性模型估计的方差约为 $\hat{\mu}+\hat{\mu}^2$。由此可见,尽管两个模型的拟合值相似,负二项广义线性模型中 $\hat{\beta}$ 的标准误更大,估计方差也更大,这反映了泊松广义线性模型所忽略的过度离散问题。

4.3.5 关于比率的泊松回归

当某种类型的事件在一定时间、空间或其他的规模范围内发生时,对事件发生的*比率*(*rate*)进行模型分析通常比分析事件的数量本身更有意义。例如,在研究某年某一城市发生的凶杀事件时,可以构建有关凶杀率的模型。将每个城市的凶杀率定义为该年所

发生的凶杀事件数除以该城市的人口规模。对凶杀率建立模型可以描述凶杀率如何受城市失业率、居民收入中位数以及居民中高中毕业的比例等因素的影响。在第 9.7 节中，我们将讨论因变量为比率的泊松回归模型。

4.3.6 关于 $I\times J$ 列联表独立性的泊松广义线性模型

泊松对数线性模型的应用之一是对列联表中的计数进行模型分析。我们以一个二维表格的独立计数 $\{Y_{ij}\}$ 为例，它们服从均值为 $\{\mu_{ij}\}$ 的泊松分布。假定 $\{\mu_{ij}\}$ 满足

$$\mu_{ij} = \mu\alpha_i\beta_j,$$

其中 $\{\alpha_i\}$ 和 $\{\beta_j\}$ 是满足条件 $\sum_i\alpha_i = \sum_j\beta_j = 1$ 的大于零的常数。这个模型的乘积形式可以通过对数连结表示为包括线性预测项的广义线性模型，

$$\log\mu_{ij} = \lambda + \alpha_i^* + \beta_j^*, \tag{4.13}$$

其中 $\lambda = \log\mu, \alpha_i^* = \log\alpha_i, \beta_j^* = \log\beta_j$。该泊松对数线性模型包括两个分类变量的主效应，但不包括交互效应。

由于 $\{Y_{ij}\}$ 是独立的，总的样本规模 $\sum_i\sum_j Y_{ij}$ 服从均值为 $\sum_i\sum_j\mu_{ij} = \mu$ 的泊松分布。以 $\sum_i\sum_j Y_{ij} = n$ 为条件，单元格计数服从概率为 $\{\pi_{ij} = \mu_{ij}/\mu = \alpha_i\beta_j\}$ 的多项分布。类似地，读者可以自行验证在以 n 为条件的情况下，行总计 $\{Y_{i+}\}$ 服从概率为 $\{\pi_{i+} = \alpha_i\}$ 的多项分布，列总计 $\{Y_{+j}\}$ 服从概率为 $\{\pi_{+j} = \beta_j\}$ 的多项分布。

以 n 为条件，上述模型是满足 $\pi_{ij} = \alpha_i\beta_j = \pi_{i+}\pi_{+j}$ 的多项分布模型。它等同于两个分类变量相互独立的情形。事实上，在泊松广义线性模型中，独立性对应的是对数线性模型（式 4.13）。第 3 章所介绍的二维表格独立性的统计推断与广义线性模型也存在着联系，它对应于泊松对数线性模型或者给定 n（或行总计，或列总计）时的多项分布模型。关于列联表分析的较复杂的对数线性模型，我们在第 8 章和第 9 章加以介绍。

4.4 广义线性模型的矩量和似然函数*

在介绍了二分数据和计数数据的广义线性模型后，接下来，我们将重点讨论相应的似然方程及其拟合方法。本章余下的部分比较偏重技术细节，主要推导可适用于后面介绍的大多数模型的一般性结果。有关进一步的讨论，参见 McCullagh 和 Nelder(1989)。

将广义线性模型的表述加以扩展，使其能够处理许多具有两个参数的分布是非常有意义的。广义线性模型的随机部分设定，关于 Y 的 N 个观测值 $(y_1,\cdots,y_N)$ 是相互独立的，y_i 的概率密度函数为

$$f(y_i;\theta_i,\phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)\}。\tag{4.14}$$

这被称为*指数离散族*（*exponential dispersion family*），其中 ϕ 为*离散参数*（*dispersion parameter*）(Jorgensen 1987)。参数 θ_i 为*自然参数*（*natural parameter*）。

当 ϕ 已知时，式 4.14 简化为自然指数族（式 4.1）的形式，即

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_iQ(\theta_i)]。$$

其中 $Q(\theta)$ 等于式 4.14 中的 $\theta/a(\phi)$，$a(\theta)$ 等于式 4.14 中的 $\exp[-b(\theta)/a(\phi)]$，$b(y)$ 等于式 4.14 中的 $\exp[c(y,\phi)]$。对于如二项分布和泊松分布这样的单一参数分布，没有必要使用一般性公式（式 4.14）。通常情况下，$a(\phi)$ 具有 $a(\phi) = \phi/\omega_i$ 的形式，其中 ω_i 是已知的权数。例如，当 y_i 是 n_i 次独立测量的均值时，如 n_i 次伯努利试验的样本比例，$\omega_i = n_i$（第 4.4.2 节）。

4.4.1 随机部分的均值和方差函数

利用公式 4.14 中的项，可以得出关于 $E(Y_i)$ 和 $\text{var}(Y_i)$ 的一般表述。令 $L_i = \log f(y_i;\theta_i,\phi)$ 表示关于 y_i 的对数似然函数，也即，对数似然函数为 $L=\sum_i L_i$，那么，由式 4.14 可得，

$$L_i = [y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)。\tag{4.15}$$

因此，

$$\partial L_i/\partial\theta_i = [y_i - b'(\theta_i)]/a(\phi),\quad \partial^2 L_i/\partial\theta_i^2 = -b''(\theta_i)/a(\phi),$$

其中 $b'(\theta_i)$ 和 $b''(\theta_i)$ 表示 $b(\cdot)$ 的前两阶导数在 θ_i 处的值。我们现在应用一般似然函数的结果

$$E\left(\frac{\partial L}{\partial\theta}\right)=0 \quad 和 \quad -E\left(\frac{\partial^2 L}{\partial\theta^2}\right)=E\left(\frac{\partial L}{\partial\theta}\right)^2,$$

它们在指数族分布所满足的规则条件下成立（Cox and Hinkley，1974，Sec. 4.8）。在只包括一个观测值的情况下，由第一个公式可得，$E[Y_i - b'(\theta_i)]/a(\phi)=0$，或者

$$\mu_i = E(Y_i) = b'(\theta_i)。\tag{4.16}$$

由第二个公式可得，

$$b''(\theta_i)/a(\phi) = E[(Y_i - b'(\theta_i)/a(\phi)]^2 = \text{var}(Y_i)/[a(\phi)]^2,$$

因此，

$$\text{var}(Y_i) = b''(\theta_i)a(\phi)。\tag{4.17}$$

综上所述，公式 4.14 中的函数 $b(\cdot)$ 决定了 Y_i 的矩量。

4.4.2 泊松分布和二项分布的均值与方差

现在，我们详细介绍泊松分布和二项分布的均值与方差。当 Y_i 服从泊松分布时，

$$f(y_i;\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp(y_i\log\mu_i - \mu_i - \log y_i!)$$
$$= \exp[y_i\theta_i - \exp(\theta_i) - \log y_i!],$$

其中 $\theta_i=\log\mu_i$。令 $b(\theta_i)=\exp(\theta_i)$，$a(\phi)=1$，以及 $c(y_i,\phi)=-\log y_i!$，该分布具有公式 4.14 的指数离散形式，它的自然参数为 $\theta_i=\log\mu_i$。由式 4.16 和式 4.17 可得，

$$E(Y_i) = b'(\theta_i) = \exp(\theta_i) = \mu_i,$$
$$\text{var}(Y_i) = b''(\theta_i) = \exp(\theta_i) = \mu_i。$$

接下来，假定 n_iY_i 服从 $\text{bin}(n_i,\pi_i)$ 分布，也即，这里 y_i 是指成功试验在样本中所占的*比例*（而不是*次数*），所以 $E(Y_i)$ 是独立于 n_i 的。令 $\theta_i=\log[\pi_i/(1-\pi_i)]$，则有 $\pi_i=\exp(\theta_i)/[1+\exp(\theta_i)]$ 以及 $\log(1-\pi_i)=-\log[1+\exp(\theta_i)]$。对公式 4.3 加以扩展，读者可以证明：

$$f(y_i;\pi_i,n_i) = \binom{n_i}{n_iy_i}\pi_i^{n_iy_i}(1-\pi_i)^{n_i-n_iy_i}$$
$$= \exp\left[\frac{y_i\theta_i - \log[1+\exp(\theta_i)]}{1/n_i} + \log\binom{n_i}{n_iy_i}\right]。\tag{4.18}$$

它具有公式 4.14 的指数离散形式，其中 $b(\theta_i)=\log[1+\exp(\theta_i)]$，$a(\phi)=1/n_i$，以及 $c(y_i,\phi)=\log\binom{n_i}{n_iy_i}$。它的自然参数为 Logit，即 $\theta_i=\log[\pi_i/(1-\pi_i)]$。由式 4.16 和式

4.17可得，

$$E(Y_i) = b'(\theta_i) = \exp(\theta_i)/[1+\exp(\theta_i)] = \pi_i,$$
$$\mathrm{var}(Y_i) = b''(\theta_i)a(\phi) = \exp(\theta_i)/\{[1+\exp(\theta_i)]^2 n_i\} = \pi_i(1-\pi_i)/n_i。$$

4.4.3 系统部分和连结函数

令$(x_{i1},\cdots,x_{ip})$表示关于第 i 个观测值的各解释变量的取值。广义线性模型的系统部分利用线性预测项将参数$\{\eta_i\}$与这些变量联系在一起，则有

$$\eta_i = \sum_j \beta_j x_{ij}, \quad i = 1,\cdots,N。$$

上式用矩阵形式表示为

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

其中 $\boldsymbol{\eta}=(\eta_1,\cdots,\eta_N)'$，$\boldsymbol{\beta}=(\beta_1,\cdots,\beta_N)'$是关于模型参数的列向量，$\mathbf{X}$ 是 $N\times p$ 矩阵，代表 N 个对象的解释变量取值。在普通线性模型中，$\mathbf{X}$ 被称为*设计矩阵*(*design matrix*)。然而，它并不一定是针对实验设计而言的，在广义线性模型的文献中，$\mathbf{X}$ 被称为*模型矩阵*(*model matrix*)。

广义线性模型通过连结函数 $g(\cdot)$将 η_i 和 $\mu_i=E(Y_i)$联系起来。因此，μ_i 与解释变量的关系为：

$$\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}, \quad i = 1,\cdots,N。$$

在式 4.14 中满足 $g(\mu_i)=\theta_i$ 的连结函数 g 为*典型连结*(*canonical link*)。此时，自然参数和线性预测项之间存在简单的关系：

$$\theta_i = \sum_j \beta_j x_{ij}。$$

由于$\mu_i=b'(\theta_i)$，自然参数是均值的函数，也即 $\theta_i=(b')^{-1}(\mu_i)$，其中$(b')^{-1}(\cdot)$表示 b'的反函数。因此，典型连结函数是 b' 的反函数。例如，在泊松模型中，$b(\theta_i)=\exp(\theta_i)$，所以 $b'(\theta_i)=\exp(\theta_i)=\mu_i$。这时，$(b')^{-1}(\cdot)$是幂函数的反函数，也就是对数函数(即，$\theta_i=\log\mu_i$)。对数连结函数是泊松模型的典型连结。

4.4.4 广义线性模型的似然方程

对于 N 个独立的观测值的广义线性模型，由式 4.15 可得，其对数似然函数为

$$L(\boldsymbol{\beta}) = \sum_i L_i = \sum_i \log f(y_i;\theta_i,\phi) = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_i c(y_i,\phi)。\quad (4.19)$$

$L(\boldsymbol{\beta})$表明，$\boldsymbol{\theta}$ 取决于模型参数 $\boldsymbol{\beta}$。

对于所有 j，相应的似然方程为

$$\partial L(\boldsymbol{\beta})/\partial\beta_j = \sum_i \partial L_i/\partial\beta_j = 0。$$

利用链式法则(chain rule)对对数似然函数(式 4.19)求导，

$$\frac{\partial L_i}{\partial\beta_j} = \frac{\partial L_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}。\quad (4.20)$$

由于$\partial L_i/\partial\theta_i=[y_i-b'(\theta_i)]/a(\phi)$，$\mu_i=b'(\theta_i)$和 $\mathrm{var}(Y_i)=b''(\theta_i)a(\phi)$(见式 4.16 和式 4.17)，所以有

$$\partial L_i/\partial\theta_i = (y_i-\mu_i)/a(\phi), \quad \partial\mu_i/\partial\theta_i = b''(\theta_i) = \mathrm{var}(Y_i)/a(\phi)。$$

同时，由于 $\eta_i = \sum_j \beta_j x_{ij}$，可得

$$\partial\eta_i/\partial\beta_j = x_{ij}。$$

最后，根据 $\eta_i = g(\mu_i)$，$\partial\mu_i/\partial\eta_i$ 取决于模型的连结函数。将这些结果代入公式 4.20，则有

$$\frac{\partial L_i}{\partial\beta_j} = \frac{y_i - \mu_i}{a(\phi)}\frac{a(\phi)}{\mathrm{var}(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}。 \tag{4.21}$$

似然方程为

$$\sum_{i=1}^{N}\frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)}\frac{\partial\mu_i}{\partial\eta_i} = 0, \quad j = 1,\cdots,p。 \tag{4.22}$$

尽管 $\boldsymbol{\beta}$ 没有直接出现在这些方程式里，但是由于 $\mu_i = g^{-1}(\sum_j \beta_j x_{ij})$，它通过 μ_i 而被间接包括进来。不同的连结函数对应于不同的方程组。

有趣的是，似然方程（式 4.22）仅通过 μ_i 和 $\mathrm{var}(Y_i)$ 取决于 Y_i 的分布。方差本身又通过某个特定的函数形式由均值所决定

$$\mathrm{var}(Y_i) = v(\mu_i),$$

以泊松分布为例，$v(\mu_i) = \mu_i$，对于伯努利分布而言，$v(\mu_i) = \mu_i(1 - \mu_i)$，对于正态分布则有 $v(\mu_i) = \sigma^2$（即同方差）。当 Y_i 服从自然指数族分布时，均值和方差之间的关系决定了这个分布的特点（Jørgensen，1987）。例如，如果 Y_i 服从自然指数分布并且 $v(\mu_i) = \mu_i$，那么 Y_i 必然服从泊松分布。

4.4.5 二项分布广义线性模型的似然方程

利用第 4.4.2 节的表示符号，假定 n_iY_i 服从 $\mathrm{bin}(n_i,\pi_i)$ 分布。这时，y_i 是在 n_i 次试验中成功的样本比例。单一预测变量的二项分布广义线性模型式（4.8）可扩展到包括多个预测变量的情况：

$$\pi_i = \Phi(\sum_j \beta_j x_{ij}), \tag{4.23}$$

其中 Φ 表示某种连续分布的标准累积分布函数（cdf）。由于 $\pi_i = \mu_i = \Phi(\eta_i)$ 以及 $\eta_i = \sum_j \beta_j x_{ij}$，

$$\partial\mu_i/\partial\eta_i = \phi(\eta_i) = \phi(\sum_j \beta_j x_{ij}),$$

其中 $\phi(u) = \partial\Phi(u)/\partial u$（即，累积分布函数 Φ 所对应的概率密度函数）。由于 $\mathrm{var}(Y_i) = \pi_i(1 - \pi_i)/n_i$，似然方程（式 4.22）简化为

$$\sum_i \frac{n_i(y_i - \pi_i)x_{ij}}{\pi_i(1 - \pi_i)}\phi(\sum_j \beta_j x_{ij}) = 0, \tag{4.24}$$

其中 $\pi_i = \Phi(\sum \beta_j x_{ij})$。似然方程取决于连结函数 Φ^{-1} 的反函数的导数。

对于 Logit 连结，$\eta_i = \log[\pi_i/(1 - \pi_i)]$，因此 $\partial\eta_i/\partial\pi_i = 1/[\pi_i(1 - \pi_i)]$，$\partial\mu_i/\partial\eta_i = \partial\pi_i/\partial\eta_i = \pi_i(1 - \pi_i)$，那么似然方程（式 4.22 和式 4.24）简化为

$$\sum_i n_i(y_i - \pi_i)x_{ij} = 0, \tag{4.25}$$

其中 π_i 对应于在公式 4.23 中 Φ 为标准 logistic 累积分布函数的情况。

4.4.6 模型参数估计值的渐近协方差矩阵

广义线性模型的似然函数也决定了最大似然估计 $\hat{\boldsymbol{\beta}}$ 的渐近协方差矩阵。这个矩阵是信息矩阵（information matrix）$\mathfrak{I}$ 的逆矩阵，而信息矩阵的元素为 $E[-\partial^2 L(\boldsymbol{\beta})/\partial\beta_h\partial\beta_j]$。为推导此结果，我们利用如下公式（Cox and Hinkley，1974，Sec. 4.8），即在指数族分布中，

对数似然函数中的 L_i 存在，

$$E\left(\frac{\partial^2 L_i}{\partial\beta_h\partial\beta_j}\right) = -E\left(\frac{\partial L_i}{\partial\beta_h}\right)\left(\frac{\partial L_i}{\partial\beta_j}\right)。$$

这样，由式 4.21 可得：

$$E\left(\frac{\partial^2 L_i}{\partial\beta_h\partial\beta_j}\right) = -E\left[\frac{(Y_i-\mu_i)x_{ih}}{\mathrm{var}(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\frac{(Y_i-\mu_i)x_{ij}}{\mathrm{var}(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right]$$
$$= \frac{-x_{ih}x_{ij}}{\mathrm{var}(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2。$$

由于 $L(\boldsymbol{\beta}) = \sum_i L_i$，所以

$$E\left(-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_h\partial\beta_j}\right) = \sum_{i=1}^{N}\frac{x_{ih}x_{ij}}{\mathrm{var}(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2。$$

从单个元素一般化到整个矩阵，信息矩阵的形式为

$$\mathfrak{I} = \mathbf{X}'\mathbf{W}\mathbf{X}, \tag{4.26}$$

其中 $\mathbf{W}$ 是对角矩阵，其主对角线上的元素为

$$w_i = (\partial\mu_i/\partial\eta_i)^2/\mathrm{var}(Y_i)。 \tag{4.27}$$

$\hat{\boldsymbol{\beta}}$ 的渐近协方差矩阵可通过

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = \hat{\mathfrak{I}}^{-1} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1} \tag{4.28}$$

来估计，其中 $\hat{\mathbf{W}}$ 是 $\mathbf{W}$ 在 $\hat{\boldsymbol{\beta}}$ 处的取值。由式 4.27 可知，$\mathbf{W}$ 的形式取决于连结函数。接下来，我们将以泊松广义线性模型为例介绍协方差矩阵，有关二项分布广义线性模型的情况可参见第 5.5 节。

4.4.7 泊松对数线性模型的似然方程和协方差矩阵

对于一般的泊松对数线性模型（式 4.4），其矩阵表述形式为

$$\log\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}。$$

在对数连结中，$\eta_i = \log\mu_i$，则有 $\mu_i = \exp(\eta_i)$ 以及 $\partial\mu_i/\partial\eta_i = \exp(\eta_i) = \mu_i$。由于 $\mathrm{var}(Y_i) = \mu_i$，似然方程（式 4.22）简化为：

$$\sum_i (y_i - \mu_i)x_{ij} = 0。 \tag{4.29}$$

这些方程中，$\boldsymbol{\beta}$ 的充分统计量 $\sum_i y_i x_{ij}$ 等于它们的期望值。同时，由于

$$w_i = (\partial\mu_i/\partial\eta_i)^2/\mathrm{var}(Y_i) = \mu_i$$

关于 $\hat{\boldsymbol{\beta}}$ 的协方差矩阵（式 4.28）的估计值是 $(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}$，其中 $\hat{\mathbf{W}}$ 是主对角线元素为 $\hat{\boldsymbol{\mu}}$ 的对角矩阵。

4.5 广义线性模型的统计推断

对于大多数广义线性模型来说，似然方程式 4.22 是 $\boldsymbol{\beta}$ 的非线性函数。在这里，我们暂不讨论如何求解方程组以得出最大似然估计值 $\hat{\boldsymbol{\beta}}$，而是主要介绍如何利用模型拟合结果进行统计推断。

在第 1.3.3 节中，我们提到了关于显著性检验和区间估计的沃尔德、计分和似然比方法，这些方法对所有广义线性模型都适用。在本节中，我们将分析广义线性模型的*偏*

离度(*deviance*),重点讨论基于似然比的统计推断。

4.5.1 偏离度和拟合优度

由第4.1.5节可得,*饱和*(*saturated*)广义线性模型中每一个观测值都对应着一个参数。饱和模型给出的是完全拟合,这听起来不错,但它本身并没有实际价值。饱和模型并不对数据进行修匀,也不具备简单模型所具有简约性。但是,如在检查其他模型的拟合状况时,可以用饱和模型作为基准。

饱和模型解释了模型系统部分的所有变动。令 $\tilde{\theta}$ 表示饱和模型关于 θ 的估计值,对所有的 i,都有 $\tilde{\mu}_i = y_i$。在某一特定的非饱和模型中,记相应的最大似然估计为$\hat{\theta}$和 $\hat{\mu}_i$。利用该模型的最大对数似然值 $L(\hat{\boldsymbol{\mu}};\mathbf{y})$ 和饱和模型的最大对数似然值 $L(\mathbf{y};\mathbf{y})$,

$$-2\log\frac{\text{该模型的似然函数最大值}}{\text{饱和模型的似然函数最大值}} = -2[L(\hat{\boldsymbol{\mu}};\mathbf{y}) - L(\mathbf{y};\mathbf{y})]$$

可以描述非饱和模型的拟合情况。该似然比统计量检验的零假设为该模型成立,而备择假设为一个更一般的模型成立。由式4.19可得,

$$\begin{aligned}&-2[L(\hat{\boldsymbol{\mu}};\mathbf{y}) - L(\mathbf{y};\mathbf{y})]\\&= 2\sum_i [y_i\tilde{\theta}_i - b(\tilde{\theta}_i)]/a(\phi) - 2\sum_i [y_i\hat{\theta}_i - b(\hat{\theta}_i)]/a(\phi)。\end{aligned}$$

通常情况下,公式4.14中的 $a(\phi)$ 具有 $a(\phi) = \phi/\omega_i$ 的形式,这时,上述统计量等于

$$2\sum_i \omega_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/\phi = D(\mathbf{y};\hat{\boldsymbol{\mu}})/\phi。\qquad(4.30)$$

这也被称为*刻度化偏离度*(*scaled deviance*),而 $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ 被称为*偏离度*(*deviance*)。模型的刻度化偏离度越大,对数据拟合得越差。对于某些广义线性模型而言,刻度化偏离度近似服从卡方分布。

4.5.2 泊松模型的偏离度

按照第4.4.2节,对于泊松广义线性模型,$\hat{\theta}_i = \log\hat{\mu}_i$,$b(\hat{\theta}_i) = \exp(\hat{\theta}_i) = \hat{\mu}_i$。相似地,其饱和模型满足 $\tilde{\theta}_i = \log y_i$ 以及 $b(\tilde{\theta}_i) = y_i$。同时 $a(\phi) = 1$,所以偏离度和刻度化偏离度(式4.30)都等于

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\sum_i [y_i\log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i]。\qquad(4.31)$$

当使用对数连结的模型包括截距项时,该参数所对应的似然方程(式4.29)为 $\sum y_i = \sum\hat{\mu}_i$。这时,偏离度表达式可简化为

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\sum_i y_i\log(y_i/\hat{\mu}_i)。\qquad(4.32)$$

在二维列联表中,它简化为第3.2.1节的 G^2 统计量(式3.11),其中 y_i 由单元格计数 n_{ij} 代替,$\hat{\mu}_i$ 由独立性拟合值 $\hat{\mu}_{ij}$代替。在第14.3节我们将看到,对于单元格数量 N 给定的列联表所对应的泊松或多项分布模型,当$\{\mu_i\}$很大时,偏离度近似服从卡方分布。

4.5.3 二项分布模型的偏离度:分组数据和未分组的数据

现在考虑在$\{n_i\}$次试验中,样本比例为$\{y_i\}$的二项分布广义线性模型。由第4.4.2节可知,$\hat{\theta}_i = \log[\hat{\pi}_i/(1-\hat{\pi}_i)]$,$b(\hat{\theta}_i) = \log[1+\exp(\hat{\theta}_i)] = -\log(1-\hat{\pi}_i)$。类似地,对于饱和模型,$\tilde{\theta}_i = \log[y_i/(1-y_i)]$,$b(\tilde{\theta}_i) = -\log(1-y_i)$。同时,由于 $a(\phi) = 1/n_i$,所以 $\phi =$

1 且 $\omega_i = n_i$。其偏离度(式 4.30)等于

$$2\sum_i n_i\left\{y_i\left(\log\frac{y_i}{1-y_i}-\log\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)+\log(1-y_i)-\log(1-\hat{\pi}_i)\right\}$$

$$=2\sum_i n_i y_i \log\frac{n_i y_i}{n_i-n_i y_i}-2\sum_i n_i y_i \log\frac{n_i\hat{\pi}_i}{n_i-n_i\hat{\pi}_i}+2\sum_i n_i\log\frac{1-y_i}{1-\hat{\pi}_i}$$

$$=2\sum_i n_i y_i \log\frac{n_i y_i}{n_i\hat{\pi}_i}+2\sum_i (n_i-n_i y_i)\log\frac{n_i-n_i y_i}{n_i-n_i\hat{\pi}_i}。$$

对第 i 种试验,$n_i y_i$ 为成功的次数,$(n_i - n_i y_i)$ 为失败的次数,其中 $i = 1, \cdots, N$。因此,上述偏离度等于对 $2N$ 个成功和失败的单元格的求和,它与包括截距项的泊松对数线性模型的偏离度(式 4.32)具有相同的形式,

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\sum 观察值\times\log(观察值/拟合值)。\tag{4.33}$$

对于二项分布的结果变量,可以将数据文件设置为预测变量的每个不同取值组合对应的成功计数和失败计数的格式,就像上文描述的一样;或者也可以使数据文件中每一个对象都对应着一个伯努利 0-1 观测值。这两种情况下的偏离度是不同的。在第一种情况下,饱和模型针对预测变量的每个不同取值组合各有一个参数,而在第二种情况下饱和模型对每一个对象各有一个参数。我们将上述两种形式的数据分别称为*分组数据*(*grouped data*)和*未分组数据*(*ungrouped data*)。在分组数据的情况下,偏离度近似服从卡方分布,而这对未分组的数据并不成立(参见习题 4.22 和 5.37)。在分组数据中,当样本规模增加时,预测变量不同取值组合的数量以及饱和模型所包括的参数数量固定不变。

4.5.4 利用偏离度进行似然比模型比较

对于泊松或二项分布模型 M,$\phi = 1$,所以偏离度(式 4.30)等于

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = -2[L(\hat{\boldsymbol{\mu}};\mathbf{y}) - L(\mathbf{y};\mathbf{y})]。\tag{4.34}$$

考虑两个模型,M_0 和 M_1 的拟合值分别为 $\hat{\boldsymbol{\mu}}_0$ 和 $\hat{\boldsymbol{\mu}}_1$,其中 M_0 是 M_1 的一个特例。这时我们称模型 M_0 *嵌套*(*nested*)于 M_1。

由于 M_0 相对于 M_1 更为简单,拟合 M_0 需要的参数数量更少。然而,一个较小参数空间所对应的最大对数似然值不可能超过较大空间的最大值。因此,$L(\hat{\boldsymbol{\mu}}_0;\mathbf{y}) \leqslant L(\hat{\boldsymbol{\mu}}_1;\mathbf{y})$。而对于所有的模型,$L(\mathbf{y};\mathbf{y})$都相同,由公式(4.34)可得

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}_1) \leqslant D(\mathbf{y};\hat{\boldsymbol{\mu}}_0)。$$

即,较简单的模型具有较大的偏离度。假定模型 M_1 成立,零假设为 M_0 成立的似然比检验统计量为

$$-2[L(\hat{\boldsymbol{\mu}}_0;\mathbf{y}) - L(\hat{\boldsymbol{\mu}}_1;\mathbf{y})]$$
$$= -2[L(\hat{\boldsymbol{\mu}}_0;\mathbf{y}) - L(\mathbf{y};\mathbf{y}] - \{-2[L(\hat{\boldsymbol{\mu}}_1;\mathbf{y}) - L(\mathbf{y};\mathbf{y})]\}$$
$$= D(\mathbf{y};\hat{\boldsymbol{\mu}}_0) - D(\mathbf{y};\hat{\boldsymbol{\mu}}_1)。$$

也即,对比两个模型的似然比统计量就是它们的偏离度之差。当 M_0 的拟合比 M_1 差时,这个统计量的值较大。

事实上,由于式 4.30 中关于饱和模型的部分相互抵消,偏离度之差也具有偏离度的形式,

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}_0) - D(\mathbf{y};\hat{\boldsymbol{\mu}}_1) = 2\sum \omega_i[y_i(\hat{\theta}_{1i} - \hat{\theta}_{0i}) - b(\hat{\theta}_{1i}) + b(\hat{\theta}_{0i})]。$$

在规则条件下,这一差值近似服从卡方零分布,其自由度等于两个模型中的参数个数

之差。

对于二项分布广义线性模型和包括截距的泊松对数线性模型，由偏离度的表达式（式 4.33）可知，偏离度之差的计算应用了观察计数和两组拟合值：

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}_0)-D(\mathbf{y};\hat{\boldsymbol{\mu}}_1)=2\sum \text{观察计数}\times\log(\text{拟合值}_1/\text{拟合值}_0)。$$

对于二分变量，比较模型拟合结果的检验与数据文件是分组格式还是未分组格式无关。在分组数据和未分组数据中，相应的饱和模型是不同的，但是在构建偏离度之差时，饱和模型的对数似然函数都相互抵消掉了。

4.5.5 广义线性模型的残差

当某一总体拟合优度检验表明广义线性模型的拟合很差时，对残差进行分析会显现模型在哪里拟合得不好。一种类型的残差与偏离度有关。在式 4.30 中令 $D(\mathbf{y};\hat{\boldsymbol{\mu}})=\sum d_i$，其中，

$$d_i=2\omega_i[y_i(\tilde{\theta}_i-\hat{\theta}_i)-b(\tilde{\theta}_i)+b(\hat{\theta}_i)]。$$

这时，观测值 i 的*偏离度残差*（*deviance residual*）为

$$\sqrt{d_i}\times\text{sign}(y_i-\hat{\mu}_i),\tag{4.35}$$

另一种残差是*皮尔逊残差*（*Pearson residual*），即

$$e_i=\frac{y_i-\hat{\mu}_i}{[\widehat{\text{var}}(Y_i)]^{1/2}}。\tag{4.36}$$

例如，对于泊松广义线性模型，$\text{var}(Y_i)=\mu_i$，其皮尔逊残差为

$$e_i=(y_i-\hat{\mu}_i)/\sqrt{\hat{\mu}_i}。$$

在二维列联表中，令 y_i 等于单元格计数 n_{ij}，$\hat{\mu}_i$ 等于独立性下的拟合值 $\hat{\mu}_{ij}$，它具有式 3.12 的形式，这时，$\sum e_{ij}^2=X^2$，即皮尔逊 X^2 统计量。类似地，偏离度残差的平方和 $\sum d_{ij}=G^2$，即独立性检验的似然比统计量。

当模型成立时，皮尔逊残差和偏离度残差的变动性小于标准正态分布的情况，这是因为它们将 y_i 与拟合均值而不是真实均值进行比较（例如，公式 4.36 的分母是对 $[\text{var}(Y_i)]^{1/2}=[\text{var}(Y_i-\mu_i)]^{1/2}$ 而不是 $[\text{var}(Y_i-\hat{\mu}_i)]^{1/2}$ 的估计）。相应地，标准化残差等于普通的残差除以其渐近标准误。对于广义线性模型，原始残差向量 $\{y_i-\hat{\mu}_i\}$ 的渐近协方差矩阵为

$$\text{cov}(\mathbf{Y}-\hat{\boldsymbol{\mu}})=\text{cov}(\mathbf{Y})[\mathbf{I}\text{-}\mathbf{Hat}]。$$

这里，**I** 为单位矩阵，**Hat** 为*帽子矩阵*（*hat matrix*），

$$\mathbf{Hat}=\mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2},\tag{4.37}$$

其中 **W** 是元素为式 4.27 的对角矩阵（Pregibon，1981）。令 $\hat{h}_i$ 表示 **Hat** 中观测值 i 对应的对角线元素估计值，称之为*杠杆矩*（*leverage*）。这时，将 $y_i-\hat{\mu}_i$ 除以其估计标准误，就得到了标准化皮尔逊残差

$$r_i=\frac{y_i-\hat{\mu}_i}{\{[\widehat{\text{var}}(Y_i)](1-\hat{h}_i)\}^{1/2}}=\frac{e_i}{\sqrt{1-\hat{h}_i}}。\tag{4.38}$$

例如，在泊松广义线性模型中，$r_i=(y_i-\hat{\mu}_i)/\sqrt{\hat{\mu}_i(1-\hat{h}_i)}$。Pierce 和 Schafer（1986）介绍了标准化偏离度残差。

在线性模型中，帽子矩阵之所以被这样命名，是因为 **Hat** × **y** 给出了数据的拟合值

$\hat{\boldsymbol{\mu}}$。对于广义线性模型，将所估计的帽子矩阵乘以关于 $g(\mathbf{y})$ 的线性近似，可得 $\hat{\boldsymbol{\eta}}=g(\hat{\boldsymbol{\mu}})$，即模型线性预测项的估计值。一个观测值的杠杆矩越大，它对模型拟合的潜在影响力也就越大。与普通回归一致，杠杆矩落在 0 和 1 之间，且它们的和等于模型参数的个数。与普通回归不同的是，杠杆矩的值既取决于拟合情况，也取决于模型矩阵，对应极端预测值的点其杠杆矩并不一定很大。

4.6 广义线性模型的拟合

终于，我们要讨论如何求解广义线性模型中参数的最大似然估计值 $\hat{\boldsymbol{\beta}}$。似然方程式 4.22 通常是 $\hat{\boldsymbol{\beta}}$ 的非线性方程。接下来，我们介绍一种求解非线性方程的通用迭代方法，以及两种与之相关的确定似然函数最大值的方法。

4.6.1 Newton-Raphson 法

Newton-Raphson 法（*Newton-Raphson method*）是一种求解非线性方程的迭代方法，由此得到的方程的解决定函数在哪个点上取最大值。这种迭代方法以对方程的解的一个初始猜测为起点。在初始猜测值附近通过二阶多项式来近似需要最大化的函数，从而求得使该多项式取最大值的点作为第二次猜测值。接下来，继续用二阶多项式来近似第二次猜测值附近的函数，使该多项式取最大值的点即为第三次猜测值。按照这种方式，可以生成一系列的猜测值。当函数具有较好特性和/或初始猜测值足够好时，这些猜测值收敛于相应函数取最大值的点。

接下来，我们具体介绍如何用 Newton-Raphson 法求解使函数 $L(\boldsymbol{\beta})$ 取最大值的点 $\hat{\boldsymbol{\beta}}$。令 $\mathbf{u}'=(\partial L(\boldsymbol{\beta})/\beta_1,\partial L(\boldsymbol{\beta})/\beta_2,\cdots)$。令 $\mathbf{H}$ 表示元素为 $h_{ab}=\partial^2 L(\boldsymbol{\beta})/\partial\beta_a\partial\beta_b$ 的矩阵，也称为*海塞矩阵*（*Hessian matrix*）。令 $\mathbf{u}^{(t)}$ 和 $\mathbf{H}^{(t)}$ 分别表示 $\mathbf{u}$ 和 $\mathbf{H}$ 在第 t 个对 $\hat{\boldsymbol{\beta}}$ 的猜测值 $\boldsymbol{\beta}^{(t)}$ 处的取值。在迭代过程（$t=0,1,2,\cdots$）的第 t 步中，用二阶泰勒级数展开式来近似 $\boldsymbol{\beta}^{(t)}$ 附近的 $L(\boldsymbol{\beta})$，

$$L(\boldsymbol{\beta})\approx L(\boldsymbol{\beta}^{(t)})+\mathbf{u}^{(t)'}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})+\left(\frac{1}{2}\right)(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})'\mathbf{H}^{(t)}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})。$$

求解 $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\approx\mathbf{u}^{(t)}+\mathbf{H}^{(t)}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})=0$，所得出的 $\boldsymbol{\beta}$ 为下一次的猜测值。这可以表示为

$$\boldsymbol{\beta}^{(t+1)}=\boldsymbol{\beta}^{(t)}-(\mathbf{H}^{(t)})^{-1}\mathbf{u}^{(t)},\tag{4.39}$$

其中，假定 $\mathbf{H}^{(t)}$ 是非奇异矩阵（但在实际运算中，可以通过标准算法来求解线性方程组，而不是直接计算逆矩阵）。

迭代过程一直进行，直到在相邻的步骤间 $L(\boldsymbol{\beta}^{(t)})$ 的变化足够小为止。最大似然估计值是当 $t\to\infty$ 时 $\boldsymbol{\beta}^{(t)}$ 的极限；然而，如果 $L(\boldsymbol{\beta})$ 存在其他的局部极值使 $L(\boldsymbol{\beta})$ 的导数等于零，上述结果不一定会出现。在这种情况下，选取一个好的初始猜测值尤为关键。计算当 $\boldsymbol{\beta}$ 只包括一个元素时每一步的拟合过程（习题 4.34），将有助于理解 Newton-Raphson 法。另外，图 4.6 演示了该方法的完整步骤，并给出了其中某一步时函数的抛物线（二阶）近似。

在下一章中，我们将通过 Newton-Raphson 法求解 logistic 回归模型。在这里，我们先用一个简单的、答案已知的例子（即最大化一个服从 $\text{bin}(n,\pi)$ 分布的观测值 y 的对数似然函数）来简要说明。由第 1.3.2 节可知，$L(\pi)=y\log\pi+(n-y)\log(1-\pi)$ 的前两阶

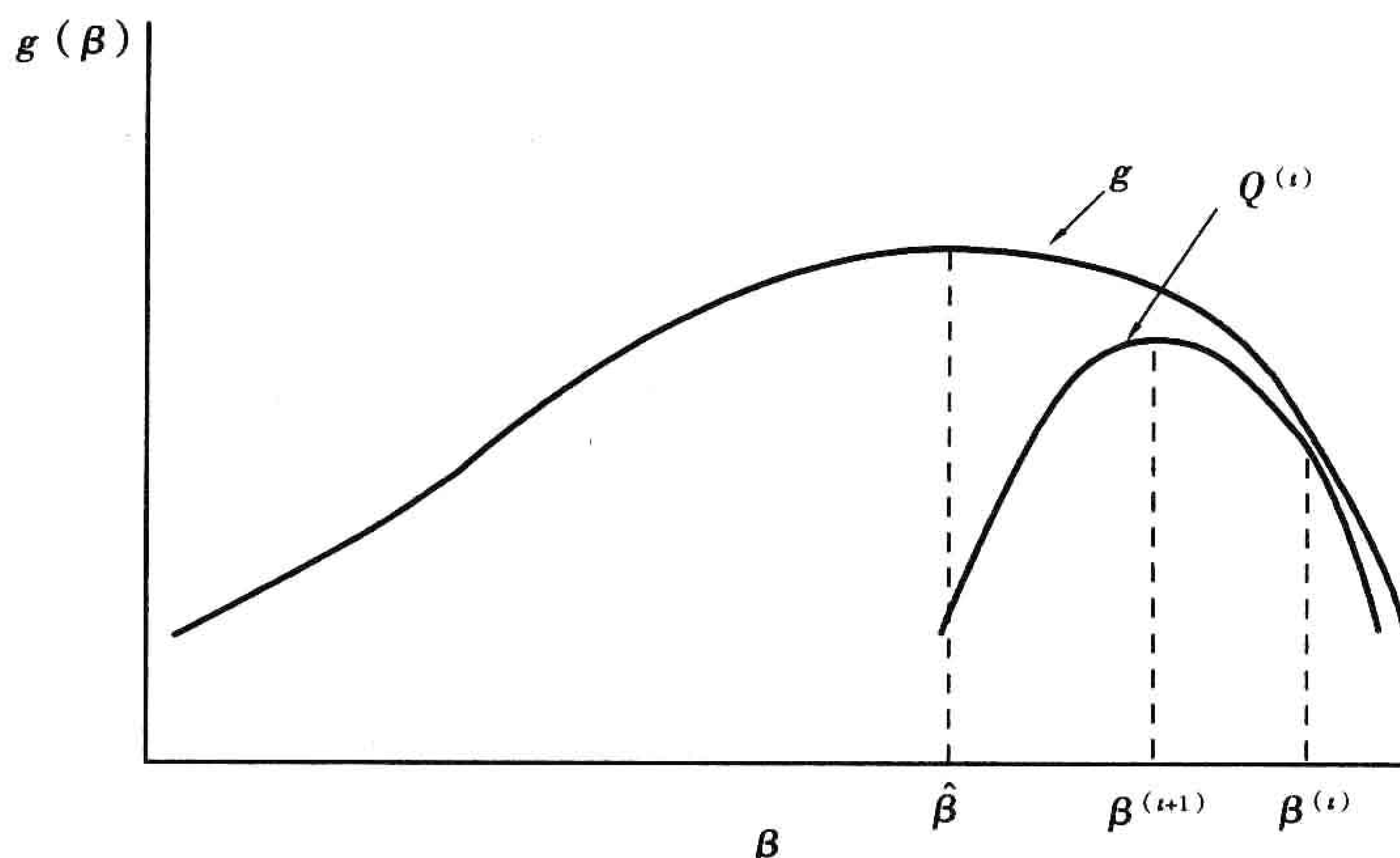


图 4.6 Newton-Raphson 法的过程

导数分别为

$$u=\frac{y-n\pi}{\pi(1-\pi)},\quad H=-\left[\frac{y}{\pi^2}+\frac{n-y}{(1-\pi)^2}\right]。$$

Newton-Raphson 法的每一步具有以下形式：

$$\pi^{(t+1)}=\pi^{(t)}+\left[\frac{y}{(\pi^{(t)})^2}+\frac{n-y}{(1-\pi^{(t)})^2}\right]^{-1}\frac{y-n\pi^{(t)}}{\pi^{(t)}(1-\pi^{(t)})}。$$

当 $y/n>\pi^{(t)}$ 时，$\pi^{(t)}$ 单调递增；当 $y/n<\pi^{(t)}$ 时，$\pi^{(t)}$ 单调递减。例如，读者可以验证，对于 $\pi^{(0)}=\frac{1}{2}$，$\pi^{(1)}=y/n$。当 $\pi^{(t)}=y/n$ 时，不需要再进行调整，并且有 $\pi^{(t+1)}=y/n$，这就是关于 $\hat{\pi}$ 的正解。对于初始值不是 $\frac{1}{2}$ 的情况，通常需要进行四到五次迭代，才能实现充分收敛。

在运用 Newton-raphson 法时，$\boldsymbol{\beta}^{(t)}$ 向 $\hat{\boldsymbol{\beta}}$ 的收敛通常都很快。在 t 足够大的情况下，对于所有的 j，满足：

$$存在某个\ c>0,\quad |\beta_j^{(t+1)}-\hat{\beta}_j|\leq c|\beta_j^{(t)}-\hat{\beta}_j|^2。$$

这被称为二阶（*second-order*）收敛。它意味着，在进行足够多次的迭代后，近似过程中正确的小数点位数大约翻倍。在实际应用中，该方法往往只需要较少的迭代次数就可以得到令人满意的收敛结果。

4.6.2 Fisher 计分法

Fisher 计分法（*Fisher scoring*）是求解似然方程的另一种迭代方法。它与 Newton-Raphson 法相似，二者的主要区别在于海塞矩阵。Fisher 计分法使用的是这个矩阵的*期望值*（*expected values*），也称为*期望信息*（*expected information*），而 Newton-Raphson 法使用的是矩阵本身，称为*观察信息*（*observed information*）。

令 $\mathfrak{I}^{(t)}$ 表示对期望信息矩阵的最大似然估计的第 t 次近似，也即，$\mathfrak{I}^{(t)}$ 的元素是 $-E(\partial^2L(\boldsymbol{\beta})/\partial\beta_a\partial\beta_b)$ 在 $\boldsymbol{\beta}^{(t)}$ 处的取值。Fisher 计分法的公式为

$$\boldsymbol{\beta}^{(t+1)}=\boldsymbol{\beta}^{(t)}+(\mathfrak{I}^{(t)})^{-1}\mathbf{u}^{(t)}$$

或者

$$\mathfrak{I}^{(t)}\boldsymbol{\beta}^{(t+1)}=\mathfrak{I}^{(t)}\boldsymbol{\beta}^{(t)}+\mathbf{u}^{(t)}。\tag{4.40}$$

对于二项分布参数的估计，由第 1.3.2 节可知，信息矩阵可简化为 $n/[\pi(1-\pi)]$。

Fisher 计分法给出

$$\pi^{(t+1)} = \pi^{(t)} + \left[\frac{n}{\pi^{(t)}(1-\pi^{(t)})}\right]^{-1} \frac{y - n\pi^{(t)}}{\pi^{(t)}(1-\pi^{(t)})}$$

$$= \pi^{(t)} + \frac{y - n\pi^{(t)}}{n} = \frac{y}{n}。$$

在完成一次迭代后,它就得出了 $\hat{\pi}$ 的答案,并在接下来的迭代过程中保持不变。

公式 4.26 给出了$\mathfrak{J} = \mathbf{X'WX}$。类似地,$\mathfrak{J}^{(t)} = \mathbf{X'W}^{(t)}\mathbf{X}$,其中 $\mathbf{W}^{(t)}$ 是 $\mathbf{W}$ 在 $\boldsymbol{\beta}^{(t)}$ 处的取值(参见公式 4.27)。这个算法同时给出,$\hat{\boldsymbol{\beta}}$ 的渐近协方差矩阵的估计值$\hat{\mathfrak{J}}^{-1}$(参见公式 4.28)等于在 t 处实现充分收敛时的$(\mathfrak{J}^{(t)})^{-1}$。由式 4.22 可得,在 Fisher 计分法和 Newton-Raphson 法中,$\mathbf{u}$ 都具有元素

$$u_j = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}。\tag{4.41}$$

在接下来的讨论中我们将看到(第 4.6.4 节),对于使用典型连结的广义线性模型,观察信息和期望信息是相同的。对于不使用典型连结的模型而言,Fisher 计分法具有以下优点:它同时能给出渐近协方差矩阵,这样期望信息矩阵必然是非负定的,并且如下文所述,它与普通线性模型的加权最小二乘法关系密切。当然,Fisher 计分法并不必然满足二阶收敛,而且对于复杂的模型而言,计算观察信息矩阵通常更容易一些。Efron 和 Hinkey(1978)在丰富 R. A. Fisher 的有关论断的同时,给出了选用观察信息的理由。他们强调,观察信息的方差估计能更好地近似相应的条件方差(以与待估计参数不相关的统计量为条件),它"更接近于数据本身",并且其结果往往与贝叶斯分析更为相近。

4.6.3 迭代再加权最小二乘法的最大似然估计*

加权最小二乘法估计(*weighted least squares estimation*)与求解最大似然估计的 Fisher 计分法存在密切联系。这里,我们讨论的是具有以下形式的广义线性模型:

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}。$$

当 $\boldsymbol{\varepsilon}$ 的协方差矩阵是 $\mathbf{V}$ 时,$\boldsymbol{\beta}$ 的加权最小二乘法(WLS)估计值为

$$(\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{z}。$$

由$\mathfrak{J} = \mathbf{X'WX}$,$\mathbf{u}$ 的元素的表达式(式 4.41),以及 $\mathbf{W}$ 的对角线元素 $w_i = (\partial\mu_i/\partial\eta_i)^2/\mathrm{var}(Y_i)$,运用公式 4.40 可得:

$$\mathfrak{J}^{(t)}\boldsymbol{\beta}^{(t)} + \mathbf{u}^{(t)} = \mathbf{X'W}^{(t)}\mathbf{z}^{(t)},$$

其中 $\mathbf{z}^{(t)}$ 的元素为

$$z_i^{(t)} = \sum_j x_{ij}\beta_j^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}$$

$$= \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}。$$

因此,Fisher 计分法的公式(式 4.40)可以表示为

$$(\mathbf{X'W}^{(t)}\mathbf{X})\boldsymbol{\beta}^{(t+1)} = \mathbf{X'W}^{(t)}\mathbf{z}^{(t)}。$$

这就是通过加权最小二乘法拟合结果变量为 $\mathbf{z}^{(t)}$ 的线性模型的普通方程组。其中,模型矩阵为 $\mathbf{X}$,协方差矩阵的逆矩阵为 $\mathbf{W}^{(t)}$。方程组的解为

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X'W}^{(t)}\mathbf{X})^{-1}\mathbf{X'W}^{(t)}\mathbf{z}^{(t)}。$$

上述公式中,向量 $\mathbf{z}$ 是连结函数 g 的线性形式在 $\mathbf{y}$ 处的取值,即

$$g(y_i) \approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i) = \eta_i + (y_i - \mu_i)(\partial\eta_i/\partial\mu_i) = z_i。\tag{4.42}$$

这个*调整后的*(或者"*操作中的*(*working*)")*结果变量* $\mathbf{z}$ 的第 i 个元素由迭代过程中第 t 步的 $z_i^{(t)}$ 来近似。该步骤以 $\mathbf{z}^{(t)}$ 为因变量对 $\mathbf{X}$ 进行加权回归来求得新的估计值 $\boldsymbol{\beta}^{(t+1)}$,其中权数为 $\mathbf{W}^{(t)}$(即,逆协方差矩阵)。这个估计值生成一个新的线性预测值 $\boldsymbol{\eta}^{(t+1)}=\mathbf{X}\boldsymbol{\beta}^{(t+1)}$ 以及用于下一次迭代的新的结果变量值 $\mathbf{z}^{(t+1)}$。对加权最小二乘法进行反复迭代可得出最大似然估计值,其中每一步迭代所用的权数矩阵都不同。这个过程被称为*迭代再加权最小二乘法*(*iterative reweighted least squares*)。

一种简单的做法是,迭代过程的第一步使用数据 $\mathbf{y}$ 作为对 $\boldsymbol{\mu}$ 的初始估计。这就确定了第一次估计的权数矩阵 $\mathbf{W}$ 以及关于 $\boldsymbol{\beta}$ 的初始估计。在第一步中,可能有必要对某些观测值进行微调以保证 $\mathbf{z}$ 的初始值——$g(\mathbf{y})$ 是有限的。例如,当 g 是计数数据对应的对数连结时,$y_i=0$ 的计数会导致问题,因此可以将其设定为 $y_i=\frac{1}{2}$。$y_i=0$ 的问题并不会影响模型本身,因为在模型中是对均值取对数,而且,在接下来的迭代过程中所拟合的均值通常都严格为正。

4.6.4 典型连结的简化*

对于使用典型连结的广义线性模型,可以进行一定的简化。对于这种连结,

$$\eta_i=\theta_i=\sum_j\beta_jx_{ij}。$$

一般来说,密度函数(式 4.14)中的 $a(\phi)$ 对于所有观测值都相等,如泊松广义线性模型($a(\phi)=1$)以及所有 $n_i=1$ 的二项分布广义线性模型(其中 $a(\phi)=1/n_i=1$)。这时,对数似然函数(式 4.19)中既包括参数又包括数据的部分是 $\sum y_i\theta_i$,它可以简化为

$$\sum_i y_i\left(\sum_j\beta_jx_{ij}\right)=\sum_j\beta_j\left(\sum_i y_ix_{ij}\right)。$$

那么,用于估计广义线性模型中的 $\boldsymbol{\beta}$ 的充分统计量是

$$\sum_i y_ix_{ij},\quad j=1,\cdots,p。$$

对于典型连结,

$$\partial\mu_i/\partial\eta_i=\partial\mu_i/\partial\theta_i=\partial b'(\theta_i)/\partial\theta_i=b''(\theta_i)。$$

因此,式 4.21 与 β_j 的似然方程有关的部分简化为

$$\frac{\partial L_i}{\partial\beta_j}=\frac{y_i-\mu_i}{\text{var}(Y_i)}b''(\theta_i)x_{ij}=\frac{(y_i-\mu_i)x_{ij}}{a(\phi)}。\tag{4.43}$$

当 $a(\phi)$ 对于所有的观测值都相等时,似然方程为

$$\sum_i x_{ij}y_i=\sum_i x_{ij}\mu_i,\quad j=1,\cdots,p。\tag{4.44}$$

这些方程式设模型参数的充分统计量等于它们的期望值(Nelder and Wedderburn,1972)。在使用恒等连结和设定正态分布的情况下,这些就是*正态方程*(*normal equations*)。我们在式 4.29 中给出了泊松对数线性模型的似然方程,并在式 4.25 中给出了二项 logistic 回归模型(当所有 $n_i=1$ 时)的似然方程。

由 $\partial L_i/\partial\beta_j$ 的表达式(式 4.43)可得,在使用典型连结的广义线性模型中,对数似然函数的二阶导数包括

$$\frac{\partial^2L_i}{\partial\beta_j\partial\beta_h}=-\frac{x_{ij}}{a(\phi)}\left(\frac{\partial\mu_i}{\partial\beta_h}\right)。$$

它不取决于观测值 y_i,所以

$$\partial^2 L(\boldsymbol{\beta})/\partial\beta_h\partial\beta_j = E[\partial^2 L(\boldsymbol{\beta})/\partial\beta_h\partial\beta_j]。$$

也即，$\mathbf{H} = -\mathfrak{I}$，就使用典型连结的模型而言，Newton-Raphson 法和 Fisher 计分法是一致的(Nelder and Wedderburn，1972)。

4.7 类似然函数与广义线性模型*

广义线性模型 $g(\mu_i) = \sum_j \beta_j x_{ij}$ 通过连结函数 g 和线性预测项来设定 μ_i。由式 4.22 和式 4.41 可知，最大似然估计 $\hat{\boldsymbol{\beta}}$ 是以下似然方程的解：

$$u_j(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{v(\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right) = 0, \quad j = 1,\cdots,p, \tag{4.45}$$

其中 $\mu_i = g^{-1}\left(\sum_j \beta_j x_{ij}\right)$，$v(\mu_i) = \text{var}(Y_i)$。这些方程式设定*计分函数*（*score functions*）$\{u_j(\boldsymbol{\beta})\}$ 等于零，$\{u_j(\boldsymbol{\beta})\}$ 也就是对数似然函数关于 $\{\beta_j\}$ 的导数。如同我们在第4.4.4 节所指出的，似然方程仅通过 μ_i 和 $v(\mu_i)$ 与 Y_i 的分布有关，而具体的分布选取决定了均值和方差之间的关系 $v(\mu_i)$。

4.7.1 均值-方差的关系决定了类似然估计

Wedderburn（1974）提出来另一种方法——*类似然估计法*（*quasi-likelihood estimation*），该方法仅假定 Y_i 的均值-方差关系，而不对具体的分布进行设定。类似然估计包括广义线性模型的连结函数和线性预测项，但与假定 Y_i 服从某种分布不同，它仅假定对于某个方差函数 v，存在

$$\text{var}(Y_i) = v(\mu_i)。$$

类似然估计的求解方程与广义线性模型的似然方程(式 4.45)相同。但是，由于未假定 $\{Y_i\}$ 服从某一自然指数分布，它们并不是似然方程。

具体而言，假定 $\{Y_i\}$ 是独立的，并且存在

$$v(\mu_i) = \mu_i。$$

类似然(QL)估计值是在式4.45中用 μ_i 替代 $v(\mu_i)$ 后的解。当进一步假定 $\{Y_i\}$ 服从指数离散分布(式 4.14)时，类似然估计值同时也是最大似然估计值。上述情况对应的是泊松分布。这时，对于 $v(\mu) = \mu$，类似然估计也是模型随机部分服从泊松分布时的最大似然估计。

Wedderburn 建议对于*所有*方差函数都使用估计方程(式 4.45)，即使其不对应于任何自然指数分布。事实上，类似然法的初衷是可以涵盖更广泛的情况，对此，我们将在第4.7.2 节加以讨论。此外，类似然估计具有与广义线性模型相同形式(式 4.28)的渐近协方差矩阵，即 $(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}$，其中 $w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(Y_i)$。

4.7.2 泊松广义线性模型的过度离散与类似然函数

前文我们已看到(第 4.3.3 节)，计数数据有时由于存在过度离散(即方差大于均值)，泊松分布假定往往不切实际，其原因之一在于对象之间的异质性。这种情况下，可以考虑使用一种不同于泊松广义线性模型的模型，该模型的均值与方差具有以下形式的关系：

$$\text{对于某一常数 } \phi, v(\mu_i) = \phi\mu_i。$$

其中，$\phi > 1$ 时对应于存在过度离散的泊松模型。

在估计方程（式 4.45）中，当 $v(\mu_i) = \phi\mu_i$ 时，ϕ 被抵消掉了。因此，该方程与泊松模型的似然方程完全一致，进而对模型参数的估计也完全相同。同时，由于

$$w_i = (\partial\mu_i/\partial\eta_i)^2 \mathrm{var}(Y_i) = (\partial\mu_i/\partial\eta_i)^2/\phi\mu_i,$$

所以该模型估计的协方差矩阵 $\mathrm{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}$ 是泊松模型的 ϕ 倍。

当方差函数具有 $v(\mu_i) = \phi v^*(\mu_i)$ 的形式时，ϕ 往往也是未知的。然而，ϕ 并不在估计方程中。令 $X^2 = \sum (y_i - \hat{\mu}_i)^2/v^*(\hat{\mu}_i)$，即 $\phi = 1$ 时的较简单模型的皮尔逊统计量。这时，X^2/ϕ 是 N 个标准化项的平方和。当 X^2/ϕ 近似于卡方分布或者当 μ_i 近似于 $\boldsymbol{\beta}$ 的线性函数且 $v^*(\hat{\mu}_i)$ 和 $v^*(\mu_i)$ 相似时，$E(X^2/\phi) \approx N - p$，即观测值数量减去模型参数的个数 p。因此，$E[X^2/(N-p)] \approx \phi$。根据矩估计的原理，Wedderburn（1974）建议使用 $\hat{\phi} = X^2/(N-p)$ 作为对协方差矩阵的倍数的估计。

总的来说，这种对计数数据的类似然法较为简单：首先拟合普通的泊松模型，并直接使用其对 p 个参数的估计值，然后再将普通模型的标准误估计乘以 $\sqrt{X^2/(N-p)}$。

现在，我们利用前面分析过的马蹄蟹数据来加以说明（见第 4.3.2 节中有关泊松广义线性模型的例子）。在使用对数连结的情况下，以壳宽来预测同伴数量的拟合为 $\log\hat{\mu} = -3.305 + 0.164x$，其中 $\hat{\beta} = 0.164$ 的标准误为 SE = 0.020。为了提高卡方统计量描述模型拟合优度的充分性，我们使用给定壳宽的雌性马蹄蟹的同伴总数及其拟合值，而不是单个雌性马蹄蟹的观察计数和拟合值。总共有 $N = 66$ 个不同的壳宽取值，每一个取值都对应着同伴的总计数 y_i 和相应拟合值 $\hat{\mu}_i$。对这些值进行比较的皮尔逊统计量为 $X^2 = 174.3$。关于标准误的类似然调整等于 $\sqrt{174.3/(66-2)} = 1.65$。因此，在这个预测方程中，关于 $\hat{\beta} = 0.164$ 的更合理的标准误为 SE = 1.65(0.020) = 0.033。

其他解决过度离散问题的方法包括允许均值在预测变量的给定取值水平上存在异质性的混合模型。就计数数据而言，这些模型包括具有随机效应的泊松广义线性模型（第 13.5 节）以及使泊松参数本身服从 γ 分布的负二项广义线性模型（第 4.3.4 节和第 13.4 节）。

4.7.3 二项分布广义线性模型的过度离散和类似然函数

类似然法也可以处理二分计数数据中的过度离散问题。当 y_i 是 n_i 个独立的二分观测值的样本均值时，其参数为 $\pi_i, i = 1, \cdots, N$，二项分布抽样具有 $E(Y_i) = \pi_i$，以及 $\mathrm{var}(Y_i) = \pi_i(1-\pi_i)/n_i$。一种简单的类似然法使用以下方差函数：

$$v(\pi_i) = \phi\pi_i(1-\pi_i)/n_i。\tag{4.46}$$

当 $\phi > 1$ 时，表明存在过度离散。由于 ϕ 在估计方程（式 4.45）中被抵消掉了，类似然估计值与二项分布模型的最大似然估计值相同。与过度离散的泊松模型一样，ϕ 是 w_i 的分母的一部分。因此，渐近协方差矩阵需要乘以 ϕ，而标准误则要乘以 $\sqrt{\phi}$。利用普通二项分布模型的 X^2 拟合统计量，对 ϕ 的估计为 $X^2/(N-p)$（Finney，1947）。

只有当模型能很好地描述 Y 的均值和预测变量之间的结构关系时，上述利用普通模型的估计值并调整其标准误的方法才是恰当的。如果模型拟合不充分，比如未包括进来一个重要的交互项，进而导致拟合优度统计量的值很大，这时对过度离散进行上述调整并不能解决相应问题。

对于二分计数数据，其他处理过度离散问题的方法包括，如具有随机效应的二项分

布广义线性模型等混合模型(第 12.3 节),以及使二项参数本身服从 β 分布的模型(第 13.3 节)。

4.7.4 例子:畸形学中的过度离散问题

表 4.5 给出了一项畸形学实验的结果,将日常饮食缺铁的受孕雌鼠分成四组。第 1 组的老鼠注射了安慰剂,其他组的老鼠则注射了补铁试剂,其中第 4 组每周都进行注射,第 2 组只在第 7 和第 10 天进行了注射,而第 3 组仅在第 0 和第 7 天进行了注射。这 58 只受孕的老鼠被观察了三周,之后对每个胎囊中死胎的数量进行了统计。在畸形学实验中,由于存在未测量的变量以及基因的差异性,在某个特定的实验组内,不同的胎囊之间死亡的概率可能存在差别。

表 4.5 在低铁畸形学研究中 58 只老鼠的结果变量计数(胎儿数量,死亡数量)

第 1 组:未补铁(低铁)
(10,1)(11,4)(12,9)(4,4)(10,10)(11,9)(9,9)(11,11)(10,10)(10,7)(12,12)
(10,9)(8,8)(11,9)(6,4)(9,7)(14,14)(12,7)(11,9)(13,8)(14,5)(10,10)
(12,10)(13,8)(10,10)(14,3)(13,13)(4,3)(8,8)(13,5)(12,12)
第 2 组:在第 7,10 天注射补铁试剂
(10,1)(3,1)(13,1)(12,0)(14,4)(9,2)(13,2)(16,1)(11,0)(4,0)(1,0)(12,0)
第 3 组:在第 0,7 天注射补铁试剂
(8,0)(11,1)(14,0)(14,1)(11,0)
第 4 组:每周注射补铁试剂
(3,0)(13,0)(9,2)(17,2)(15,0)(2,0)(14,1)(8,0)(6,0)(17,0)

来源:Moore and Tsiatis (1991)。

令 $y_{i(g)}$ 表示在实验组 g 中第 i 个胎囊所包括的 $n_{i(g)}$ 个胚胎中出现死亡的比例,并令 $\pi_{i(g)}$ 表示在该胎囊中死胎的概率。考虑关于服从 $\text{bin}(n_{i(g)},\pi_{i(g)})$ 分布的变量 $n_{i(g)}y_{i(g)}$ 的模型,其中

$$\pi_{i(g)} = \pi_g, \quad g = 1,2,3,4。$$

也即,该模型设定第 g 组中的所有胎囊具有相同的死亡概率 π_g。其最大似然估计值为 $\hat{\pi}_g$,等于该组所有胎囊中死亡胚胎的样本比例。这些估计值分别为 $\hat{\pi}_1 = 0.758(\text{SE} = 0.024)$,$\hat{\pi}_2 = 0.102(\text{SE} = 0.028)$,$\hat{\pi}_3 = 0.034(\text{SE} = 0.024)$ 以及 $\hat{\pi}_4 = 0.048(\text{SE} = 0.021)$。其中,第 g 组的标准误 $\text{SE} = \sqrt{\hat{\pi}_g(1-\hat{\pi}_g)/(\sum_i n_{i(g)})}$。由此可见,在注射安慰剂的一组中,所估计的死亡概率明显更高。

对于第 g 组中的第 i 个胎囊,拟合的死亡数为 $n_{i(g)}\hat{\pi}_g$,拟合的存活数为 $n_{i(g)}(1-\hat{\pi}_g)$。利用皮尔逊统计量将这些($N=58$)胎囊中死亡数与存活数的拟合值和观察值相比较,可得 $X^2=154.7$,自由度为 $\text{df}=58-4=54$。这里存在明显的过度离散问题。通过类似然法,$\{\hat{\pi}_g\}$ 与二项分布最大似然估计相同;但是,$\hat{\phi}=X^2/(N-p)=154.7/(58-4)=2.86$,所以相应的标准误应乘以 $\hat{\phi}^{1/2}=1.69$。

即便对过度离散进行了调整,仍然存在着明显的证据表明,注射安慰剂组的死亡概率要高得多。例如,$\pi_1-\pi_2$ 的 95% 的置信区间为

$$(0.758-0.102) \pm 1.96\left[(1.69\times 0.024)^2+(1.69\times 0.028)^2\right]^{1/2}$$

或 (0.54,0.78)。

但是,与忽略过度离散问题、用于比较相互独立的比例的沃尔德区间(0.59,0.73)相比,这个区间更大一些。

4.8 广义可加模型*

广义线性模型将普通的线性回归扩展到允许非正态分布及对均值的函数构建模型的情况。类似然函数对此做了进一步的扩展,它不需要假定某个特定分布,而只需要设定方差与均值的关系。在本节中,我们介绍另一种扩展,即利用预测变量的修匀函数来替代模型中的线性预测项。

4.8.1 修匀数据

广义线性模型的结构 $g(\mu_i) = \sum_j \beta_j x_{ij}$ 可以一般化为

$$g(\mu_i) = \sum_j s_j(x_{ij}),$$

其中 $s_j(\cdot)$ 是关于第 j 个预测变量的待定的修匀函数。其中,一个有用的修匀函数是三*次样条*(*cubic spline*)。三次样条修匀是在一组互不连结的区间中,每个区间都对应着各自的三次多项式,这些多项式在区间的边界上平滑地连结起来。

与广义线性模型一样,这个模型也需要设定随机部分的分布以及连结函数 g,由此形成的模型被称为*广义可加模型*(*generalized additive model*),简写为 GAM(Hastie and Tibshirani ,1990)。广义线性模型是当广义可加模型中的每个 s_j 都为线性函数时的特例。另外,也可以令某些 s_j 为修匀函数,而另一些为线性函数或者代表分类预测变量的虚拟变量。

拟合广义可加模型的详细介绍超出了本书的范围。简单地说,它的拟合是基于对 Newton-Raphson 算法的扩展,其中应用了局部修匀的方法。这对应于从对数似然函数中减去一个惩罚函数(penalty function),该函数会随着修匀函数变得更加波动而上升。模型拟合对可加的预测变量中的每个 s_j 设定一个偏离度以及近似的自由度,以保证可以对这些项进行统计推断。例如,一个自由度为 df = 5 的修匀函数的复杂性与一个四阶多项式相似,后者包括 5 个参数。对自由度(或修匀参数)的选取决定了相应的广义可加模型的修匀程度。

通常需要尝试不同程度的修匀,以找到一个既能充分修匀数据(以避免其趋势太不规则),又保证不过度修匀(而隐匿了有意义的模式)的模型。通过这种方法可能发现一个使用特定连结的线性模型是充分的,或者找出改进线性程度的方式。对于不包括广义可加模型的软件,可以通过加权回归的方法来修匀数据,即在预测某一点的值时,对其附近的观测值赋予较大的权数,这种*局部加权最小二乘法回归*(*locally weighted least squares regression*)一般被称为 *lowess*。不过,我们推荐使用广义可加模型,因为它能明确地识别出结果变量的形式。例如,对于二分结果变量,lowess 可能会给出小于 0 或大于 1 的预测值,广义可加模型则不会出现这种情况。

即使已经确定使用广义线性模型,广义可加模型仍会对探索性分析有帮助。例如,对于连续结果变量 Y 以及连续变量 X,散点图会给出 Y 和 X 关系的视觉信息。对于二分结果变量,如图 4.7 所示,这种图不是很有意义。由于无需假定某个特定的函数关系,绘制对预测变量所拟合的修匀函数可能会显现出某种一般趋势。

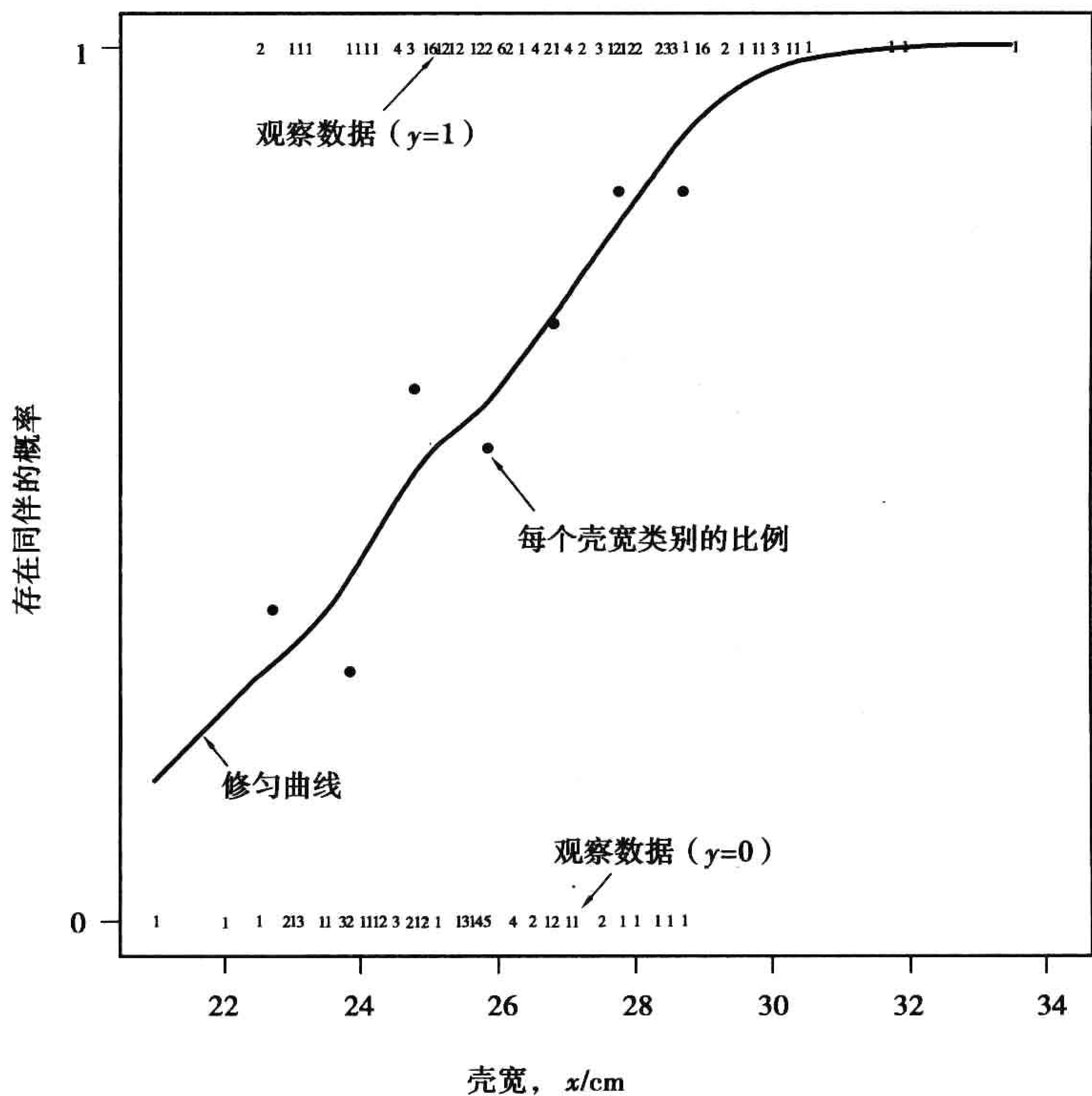


图 4.7 通过广义可加模型的修匀拟合
雌性马蹄蟹壳宽与是否存在同伴(1,是;0,否)的关系

4.8.2 关于马蹄蟹例子的广义可加模型

在第 4.3.2 节中,图 4.4 显示了关于马蹄蟹的壳宽与同伴数量关系的趋势。其中的修匀曲线是一个广义可加模型的拟合结果,该模型假定泊松分布并使用了对数连结。

在下一章,我们将通过 logistic 回归对马蹄蟹至少存在一个同伴的概率进行模型分析。对于第 i 只马蹄蟹,如果它至少拥有一个同伴,记 $y_i=1$,否则 $y_i=0$。图 4.7 展示了 y_i 和 x = 马蹄蟹的壳宽之间的关系。它包含一组 $y_i=1$ 的点以及一组 $y_i=0$ 的点。数字符号表示落在每个点上的观测值数量。从该图来看,在 x 取较大值时,$y_i=1$ 出现得更频繁。图 4.7 也给出了一条通过广义可加模型修匀数据的曲线,该模型假定二项分布并使用了 Logit 连结。这个曲线显示出一个大致上升的趋势,它比仅观察二分数据提供了更多的信息。从该曲线来看,S 型的回归函数能够较好地描述这个关系。

注 解

第 4.1 节:广义线性模型

4.1 式 4.1 所代表的分布被称为一种*自然*(或*线性*)指数分布族,以区别于在指数项中用 $r(y)$ 代替 y 的更一般的指数分布。有关其他的扩展,参见:Jørgensen(1987)。关于广义线性模型以及相关模型的书,按照难度从高到低排列,大致包括:McCullagh and Nelder(1989)、Fahrmeir and Tutz(2001)、Aitkin et al.(1989)、Dobson(2002)、Gill(2000)。另见:Firth(1991)。

第 4.3 节:计数数据的广义线性模型

4.2 有关计数数据的泊松回归以及相应模型的进一步讨论,参见:Breslow(1984)、

Cameron and Trivedi (1998)、Frome (1983)、Hinde (1982)、Lawless (1987)、Seeber (1998)，以及这些研究所提及的其他文献。

第 4.4 节：广义线性模型的矩量和似然函数

4.3 式 4.4 的函数 $b(\cdot)$ 被称为*累积函数*(*cumulant function*)，因为当 $a(\phi)=1$ 时，它的导数给出了分布的累积项(Jørgensen, 1987)。

对于许多广义线性模型，包括使用对数连结的泊松模型和使用 Logit 连结的二分模型，在满秩模型矩阵下，海塞矩阵是负定的，并且对数似然函数是严格的凹函数。这时，在相当宽泛的条件下，模型参数的最大似然估计存在并且唯一(Wedderburn, 1976)。

第 4.5 节：广义线性模型的统计推断

4.4 在 $\text{cov}(\hat{\boldsymbol{\beta}})$（参见式 4.28）、标准化皮尔逊残差的帽子矩阵（参见式 4.38）以及 Fisher 计分法（参见式 4.40）中，所使用的矩阵 **W** 是线性形式的 $g(\mathbf{y})$ 的协方差矩阵的逆矩阵（参见第 4.6.3 节）。

McCullagh 和 Nelder(1989, chap. 12)讨论了关于广义线性模型的模型诊断。有关残差的讨论，另见：Green (1984)、Pierce and Schafer (1986)、Pregibon (1980, 1981)、Williams (1987)。Pregibon(1982)证明了标准化皮尔逊残差的平方是检验观测值是否为奇异值的计分统计量。Davison 和 Hinkley(1997, Sec. 7.2)讨论了广义线性模型中的重抽样自举法(bootstrapping)。

第 4.6 节：广义线性模型的拟合

4.5 Fisher(1935b)提出了对 probit 模型进行最大似然估计的 Fisher 计分法。有关拟合广义线性模型以及迭代再加权最小二乘法与最大似然估计关系的进一步讨论，参见：Green (1984)、Jørgensen (1983)、McCullagh and Nelder (1989)、Nelder and Wedderburn(1972)。Green (1984)、Jorgensen (1983)，以及 Palmgren 和 Ekholm (1987)还讨论了在指数族非线性模型中的相应关系。

第 4.7 节：类似然函数与广义线性模型

4.6 关于类似然函数的更多讨论，参见第 11.4、第 12.6.4 和第 13.3 节，另参见：Breslow (1984)、Cox (1983)、Firth (1987)、Hinde and Demétrio (1998)、McCullagh (1983)、McCullagh and Nelder(1989)、Nelder and Pregibon(1987)、Wedderburn(1974, 1976)。相应的理论探讨，参见：Heyde(1997)。

第 4.8 节：广义可加模型

4.7 除广义可加模型外，其他的非参数修匀方法也可以描述二分结果变量和预测变量之间的关系。例如，参见：Copas(1983)、Lloyd(1999, Chap. 5)，以及第 15.3.3 节关于核修匀的介绍、Kauermann 和 Tutz(2001)关于随机效应模型的讨论。

习题

应用部分

4.1 在 2000 年美国总统选举中，佛罗里达州的棕榈滩县(Palm Beach County)成了不寻常投票模式(出现了大量无效的双重选票)的焦点，这明显是由一种令人迷惑的“蝴蝶选票(butterfly ballot)”所造成的。许多选民宣称，当他们计划投票给 Al Gore 时，错误地投给了改革党候选人 Pat Buchanan。图 4.8 显示了在佛罗里达各县 Buchanan 的总得票数以及 1996 年改革党候选人(Ross Perot)的得票数(有关细节请参见：A. Agresti

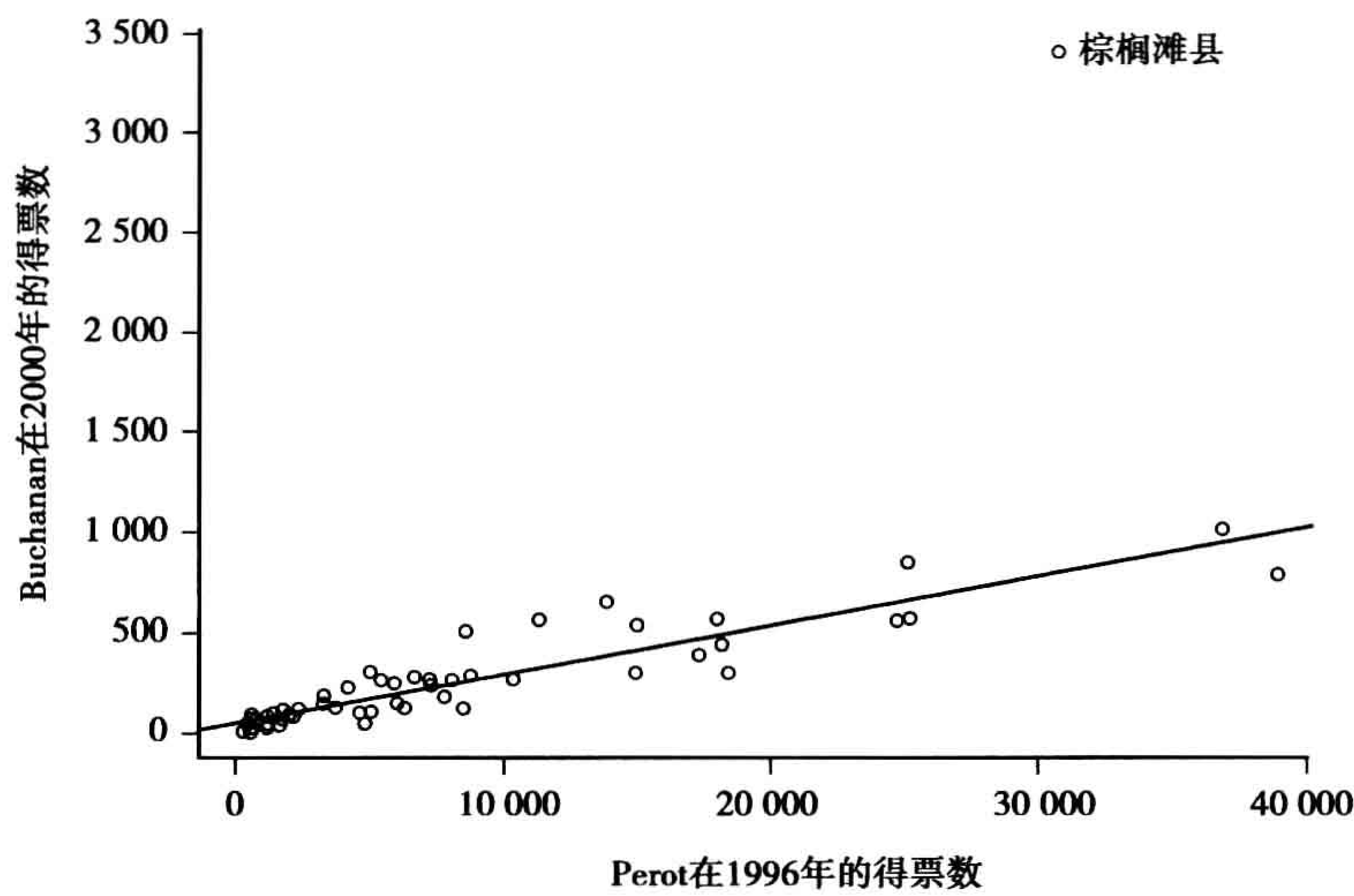


图 4.8　改革党候选人 Buchanan(2000 年)和 Perot(1996 年)在佛罗里达各县的总得票数

and B. Presnell, *J. Law Public Policy*, Volumn 13, Fall 2001, 117-134)。

a. 在第 i 个县,令 π_i 表示投票给 Buchanan 的比例,x_i 表示在 1996 年投票给 Perot 的比例。对除棕榈滩县外的所有县拟合的线性概率模型为,$\hat{\pi}_i = -0.000\ 3 + 0.030\ 4x_i$。给出下面解释中的 P 值:在 2000 年投票给 Buchanan 的估计比例大约是 1996 年投票给 Perot 的 $P\%$。

b. 对于棕榈滩县,$\pi_i = 0.007\ 9$, $x_i = 0.077\ 4$。这个结果像是一个奇异值吗?加以说明。

c. 拟合 logistic 回归,$\log[\hat{\pi}_i/(1-\hat{\pi}_i)] = -7.164 + 12.219x_i$。求棕榈滩县的 $\hat{\pi}_i$。该县对于这个模型而言是一个奇异值吗?

4.2　在美国棒球联盟 90 年间的比赛中,表 4.6 给出了首发投手完成一整场比赛的百分比。

表 4.6　习题 4.2 的数据

年　代	完成百分比	年　代	完成百分比	年　代	完成百分比
1900—1909	72.7	1930—1939	44.3	1960—1969	27.2
1910—1919	63.4	1940—1949	41.6	1970—1979	22.5
1920—1929	50.0	1950—1959	32.8	1980—1989	13.3

来源:数据取自 George Will, *Newsweek*, *Apr.* 10, 1989。

a. 将每十年发生的比赛数视为相等的,线性概率模型的最大似然拟合为 $\hat{\pi} = 0.757\ 8 - 0.069\ 4x$,其中 x = 每个十年($x = 1, 2, \cdots, 9$)。解释 0.757 8 和 −0.069 4。

b. 分别代入 $x = 10, 11, 12$,预测以后三十年完成比赛的百分比。这些预测有可能发生吗?为什么?

c. 对该数据进行 logistic 回归的最大似然拟合结果为 $\hat{\pi} = \exp(1.148 - 0.315x)/[1 + \exp(1.148 - 0.315x)]$。给出当 $x = 10, 11, 12$ 时的 $\hat{\pi}_i$。这些值发生的可能性是不是更大一些?

4.3 对表3.7中的饮酒量赋值为(0,0.5,1.5,4.0,7.0),以下是关于畸形的线性概率模型的最大似然拟合结果。

参数	估计值	标准误	沃尔德95%置信区间	
截距	0.002 5	0.000 3	0.001 9	0.003 2
饮酒量	0.001 1	0.000 7	−0.000 3	0.002 5

解释模型的拟合情况。利用该结果,估计饮酒量分别为0和7.0时出现畸形的相对风险。

4.4 在表4.2中,利用以下赋值重新拟合线性概率模型或logistic回归模型:(a)(0,2,4,6),(b)(0,1,2,3),(c)(1,2,3,4)。比较这三种情况下的$\hat{\beta}$,并对比相应的拟合值。总结对赋值进行线性转换同时保持间距的相对大小不变的影响。

4.5 在表4.3中,如果马蹄蟹至少具有一个同伴,记$Y=1$,否则令$Y=0$。利用$x=$体重拟合线性概率模型。

a. 使用普通最小二乘法。解释参数的估计值。求在体重的最大观察值(5.20千克)时的估计概率。对结果加以评论。

b. 使用最大似然估计拟合该模型,将Y视为二分变量(拟合失败是由于所拟合的概率落在了(0,1)范围外。(a)部分的拟合结果是针对正态随机部分的最大似然估计,对此拟合值落在这个范围外是允许的)。

c. 拟合logistic回归模型。证明体重为5.20千克时对应的拟合概率等于0.996 8。

d. 拟合probit模型。求体重为5.20千克时的拟合概率。

4.6 一项实验分析制作用于电脑芯片的硅片在两种生产过程中的瑕疵率。使用方法A的10个硅片中出现的瑕疵数量分别为8,7,6,6,3,4,7,2,3,4。使用方法B的10个硅片中出现的瑕疵数量分别为9,9,8,14,8,13,11,5,7,6。将这些计数视为均值为μ_A和μ_B的独立泊松变量。

a. 拟合模型$\log\mu=\alpha+\beta x$,其中$x=1$表示方法B,$x=0$表示方法A。证明$\exp(\beta)=\mu_B/\mu_A$,并解释模型的估计结果。

b. 运用沃尔德或似然比检验来检验$H_0:\beta=0$,从而检验$H_0:\mu_A=\mu_B$是否成立。对结果加以解释。

c. 构建关于μ_B/μ_A的95%的置信区间(提示:首先构建关于β的置信区间)。

d. 基于以下结果检验$H_0:\mu_A=\mu_B$:如果Y_1和Y_2分别服从均值为μ_1和μ_2的独立泊松分布,那么$(Y_1|Y_1+Y_2)$服从$n=Y_1+Y_2$和$\pi=\mu_1/(\mu_1+\mu_2)$的二项分布。

4.7 对于表4.3的数据,表4.7给出了关于$X=$体重和$Y=$同伴数量的泊松对数线性模型所拟合的SAS输出结果。

a. 估计雌蟹在体重的均值2.44千克时的$E(Y)$。

b. 通过$\hat{\beta}$描述体重的效应,说明结果中所给出的置信区间是如何构建的。

c. 构建关于Y独立于X的沃尔德检验。加以解释。

d. 你能对上述假设进行似然比检验吗?如果不行,还需要哪些信息?

e. 这里存在过度离散的问题吗?如果有必要,对标准误进行调整并加以解释。

4.8 参考习题4.7。使用恒等连结,对$X=$体重拟合可得,$\hat{\mu}=-2.60+2.264x$,其中$\hat{\beta}=2.264$的标准误为SE=0.228。重做上题(a)至(c)部分。

表 4.7 习题 4.7 的 SAS 输出结果

	Criterion		DF	Value		
	Deviance		171	560.866 4		
	Pearson Chi-Square		171	535.895 7		
	Log Likelihood			71.952 4		

Parameter	Estimate	Std Error	Wald 95%	Conf Limits	Chi-Sq	Pr > ChiSq
Intercept	−0.428 4	0.178 9	−0.779 1	−0.077 7	5.73	0.016 7
Weight	0.589 3	0.065 0	0.461 9	0.716 7	82.15	<.0 001

译者注—Weight:体重。

4.9 参见表 4.3。

a. 拟合一个泊松对数线性模型,利用 W = 体重和 C = 颜色来预测 Y = 同伴数量。将 C 视为定类变量,生成相应的虚拟变量。解释参数的估计值。

b. 估计下列颜色的雌蟹在平均体重 2.44 千克时的 $E(Y)$:(i) 颜色为浅褐色,(ii) 颜色为黑色。

c. 检验是否有必要将颜色包括在模型中(*提示*:由第 4.5.4 节可知,比较模型的似然比统计量等于它们的偏离度之差)。

d. 颜色效应的估计值在四个类别间是单调的。拟合一个将 C 当作定距变量并假定其效应为线性的简单模型。解释颜色的效应,并重做(b)和(c)部分。将该模型的拟合结果与(a)部分的模型进行比较,加以解释。

e. 将壳宽加入模型。壳宽和体重之间存在的高度正相关会有什么影响?是否有必要将两个变量都包括在模型中?

4.10 参见第 4.3.2 节中使用恒等连结的泊松模型。最小二乘法的拟合结果为 $\hat{\mu} = -10.42 + 0.51x$(SE = 0.11)。说明为什么参数的估计值不同,且标准误的值相差很大。

4.11 以壳宽为马蹄蟹同伴数的预测变量,在使用对数连结所拟合的负二项模型中,$\hat{\alpha} = -4.05$, $\hat{\beta} = 0.192$(SE = 0.048), $\hat{k}^{-1} = 1.106$(SE = 0.197)。对此结果加以解释。为什么 $\hat{\beta}$ 的标准误与第 4.3.2 节中相应泊松广义线性模型的标准误 SE = 0.020存在很大差别?哪个标准误更恰当?为什么?

4.12 参见习题 4.6。方法 A 的样本均值和方差分别为 5.0 和 4.2,方法 B 的分别为 9.0 和 8.4。

a. 将制作方法视为虚拟变量,包括此变量的泊松模型是否存在过度离散的问题?请解释。

b. 拟合负二项对数线性模型。注意,所估计的离散参数为 0,并且所估计的均值和标准误与泊松对数线性模型相同。

c. 对于所有 20 个样本观测值,样本均值和方差为 7.0 和 10.2。分别在泊松和负二项分布假设下,拟合只包括截距项的对数线性模型。比较模型结果以及总体均值的置信区间。为什么它们有差别?(*注意*:这表明,当一个重要的协变量未被测量时,泊松模型的结果会变差。)

4.13 表 4.8 给出了在 2000 年 NBA(篮球)季后赛中,洛杉矶湖人队的沙克·奥尼尔在每场比赛中的罚球情况。解说员评论说,他的罚球在每场比赛之间波动很大。对

于第 i 场比赛,假定 Y_i =“在 n_i 次罚球中命中的次数”是一个 $\text{bin}(n_i, \pi_i)$ 变量,并且 $\{Y_i\}$ 是相互独立的。

a. 拟合模型 $\pi_i = \alpha$,求 $\hat{\alpha}$ 及其标准误,并加以解释。模型看上去拟合得充分吗?(注意:你可以通过一个关于比赛和二分结果的 23×2 表格的小样本独立性检验来进行检查。)

b. 针对过度离散问题调整标准误。利用初始的和调整后的标准误,求关于 α 的 95% 的置信区间并加以比较。解释结果。

表 4.8 习题 4.13 的数据

比赛	命中数	罚球数	比赛	命中数	罚球数	比赛	命中数	罚球数
1	4	5	9	4	12	17	8	12
2	5	11	10	1	4	18	1	6
3	5	14	11	13	27	19	18	39
4	5	12	12	5	17	20	3	13
5	2	7	13	6	12	21	10	17
6	7	10	14	9	9	22	1	6
7	6	14	15	7	12	23	3	12
8	9	15	16	3	10			

来源:www.nba.com。

4.14 参见表 13.6。将种族作为虚拟变量,拟合一个对数线性模型:(a)假定服从泊松分布,(b)通过类似然法允许过度离散。比较结果。

4.15 参见习题 4.6。按照硅涂层的厚度对硅片进行了划分($z=0$,低;$z=1$,高)。每一组的前五个瑕疵计数对应于 $z=0$ 的情况,后五个对应于 $z=1$ 的情况。分析这些数据。

4.16 参见关于性行为频率的表 13.9。分析这些数据。

理论与方法

4.17 解释广义线性模型中连结函数的意义。什么是恒等连结?说明为什么它在二项分布或泊松结果变量中不常用。

4.18 当 k 已知时,证明负二项分布(式 4.12)具有自然参数为 $\log[\mu/(\mu+k)]$ 的自然指数形式(式 4.1)。

4.19 对于二分数据,定义一个使用对数连结的广义线性模型。证明其效应对应的是相对风险。你认为为什么这个连结并不常用?(提示:当线性预测项取正值时会出现什么情况?)

4.20 在 logistic 回归模型(式 4.6)中,对于 $\beta>0$ 的情况,证明:(a)随着 $x \to \infty$,$\pi(x)$ 单调增加;(b)$\pi(x)$ 的曲线是均值为 $-\alpha/\beta$ 和标准差为 $\pi/(|\beta|\sqrt{3})$ 的 logistic 分布的累积分布函数。

4.21 给出二项分布相对应的表达式(4.18)。

4.22 令 Y_i 表示关于第 i 组的 $\text{bin}(n_i, \pi_i)$ 变量,$i=1,\cdots,N$,其中 $\{Y_i\}$ 相互独立。考虑模型 $\pi_1 = \cdots = \pi_N$,用 π 来表示这个相同值。对于观测值 $\{y_i\}$,证明 $\hat{\pi} = \frac{\sum y_i}{\sum n_i}$。当所有 $n_i = 1$ 时,检验这个模型在 $N \times 2$ 表格中的拟合情况,证明 $X^2 = n$。由此可见,对

未分组的数据来说，拟合优度统计量可能毫无意义(另见习题 5.37)。

4.23 假定 Y_i 是一个 $g(\mu_i)=\alpha+\beta x_i$ 的泊松变量，其中 A 组中的 $i=1,\cdots,n_A$ 对应着 $x_i=1$，B 组中的 $i=n_A+1,\cdots,n_A+n_B$ 对应着 $x_i=0$。证明对于任意连结函数 g，似然方程(式 4.22)意味着，拟合的均值 $\hat{\mu}_A$ 和 $\hat{\mu}_B$ 等于样本均值。

4.24 对于在 n_i 次试验中样本比例为 y_i 的二分数据，我们通过类似然法拟合一个方差函数为式 4.46 的模型。证明参数估计值与二项分布广义线性模型相同，但是协方差矩阵相应乘以了 ϕ。

4.25 在连结函数的反函数为 Φ 的二项分布广义线性模型 $\pi_i=\Phi(\sum_j \beta_j x_{ij})$ 中，假定 $n_i Y_i$ 服从 $\text{bin}(n_i,\pi_i)$ 分布，求公式 4.27 中的 w_i，并求 $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$。对于 logistic 回归，证明 $w_i=n_i\pi_i(1-\pi_i)$。

4.26 一个广义线性模型中的参数 $\boldsymbol{\beta}$ 存在充分统计量 S。拟合优度检验统计量 T 的观察值为 t_o。如果 $\boldsymbol{\beta}$ 已知，检验的 P 值为 $P=P(T\geqslant t_o;\boldsymbol{\beta})$。说明为什么 $P=P(T\geqslant t_o|S)$ 是对 P 的一致最小方差无偏估计(uniform minimum variance unbiased estimator)。

4.27 令 y_{ij} 表示第 i 组中的第 j 个观测值的计数变量，$i=1,\cdots,I$，$j=1,\cdots,n_i$。假定 $\{Y_{ij}\}$ 为满足 $E(Y_{ij})=\mu_i$ 的独立泊松变量。

a. 证明关于 μ_i 的最大似然估计为 $\hat{\mu}_i=\bar{y}_i=\sum_j y_{ij}/n_i$。

b. 对这个模型的偏离度的表述进行简化(关于这个模型的检验，由 Fisher(1970，第 58 页，最早发表于 1925 年)可知，偏离度和皮尔逊统计量 $\sum_i\sum_j(y_{ij}-\bar{y}_i)^2/\bar{y}_i$ 近似服从自由度为 $\text{df}=\sum_i(n_i-1)$ 的卡方分布。在单一组的情况下，Cochran(1954)把 $\sum_j(y_{1j}-\bar{y}_1)^2/\bar{y}_1$ 称为泊松分布拟合结果的方差检验(*variance test*)，因为它将样本方差与估计的泊松方差 $\bar{y}_1$ 进行对比)。

4.28 以 λ 为条件，Y 服从均值为 λ 的泊松分布。λ 的值按照 γ 密度函数(式 13.12)变动，并有 $E(\lambda)=\mu$，$\text{var}(\lambda)=\mu^2/k$。证明 Y 的边际分布是负二项分布(式 4.12)。说明为什么负二项模型是处理过度离散的泊松模型的一种方法。

4.29 考虑式 4.8 和式 4.9 的二分模型。假定标准累积分布函数 Φ 对应于一个在 0 点两侧对称的概率密度函数 ϕ。

a. 证明在 $\pi(x)=0.5$ 时，$x=-\alpha/\beta$。

b. 证明在 $\pi(x)=0.5$ 时，$\pi(x)$ 的变动率为 $\beta\phi(0)$。在 Logit 连结的情况下，证明该变动率等于 0.25β；在 probit 连结的情况下，证明该变动率等于 $\beta/\sqrt{2\pi}$(其中 $\pi=3.14\cdots$)。

c. 证明 probit 回归曲线具有均值为 $-\alpha/\beta$ 和标准差为 $1/|\beta|$ 的正态累积分布函数的形状。

4.30 证明给定 σ 的正态分布 $N(\mu,\sigma^2)$ 满足式 4.1 的分布族，并指出其中各项的表达式。将普通回归模型表述为一个广义线性模型。

4.31 在习题 4.30 中，当 σ 也是一个参数时，证明它满足指数离散分布族(式 4.14)。

4.32 对于二分观测值，考虑模型 $\pi(x)=\dfrac{1}{2}+(1/\pi)\tan^{-1}(\alpha+\beta x)$。哪种分布具有此形式的累积分布函数？说明在什么情况下利用这个曲线的广义线性模型可能比 logistic 回归更适当。

4.33 给出下列观测值的偏离度残差(式 4.35):(a)二项分布广义线性模型,(b)泊松广义线性模型。通过满足独立性模型的二维列联表单元格计数来展示(b)部分。

4.34 考虑使函数 $L(\beta)$ 最大化的值 $\hat{\beta}$。令 $\beta^{(0)}$ 表示初始猜测值。

a. 利用 $L'(\hat{\beta}) = L'(\beta^{(0)}) + (\hat{\beta} - \beta^{(0)})L''(\beta^{(0)}) + \cdots$,说明当 $\beta^{(0)}$ 与 $\hat{\beta}$ 相近时,近似存在 $0 = L'(\beta^{(0)}) + (\hat{\beta} - \beta^{(0)})L''(\beta^{(0)})$。求解这个方程式以获得关于 $\hat{\beta}$ 的近似值 $\beta^{(1)}$。

b. 令 $\beta^{(t)}$ 表示对 $\hat{\beta}$ 的第 t 次近似值,$t = 0,1,2,\cdots$。证明下一次近似值等于

$$\beta^{(t+1)} = \beta^{(t)} - L'(\beta^{(t)})/L''(\beta^{(t)})。$$

4.35 n 个独立的观测值服从泊松分布,证明对于所有的 $t > 0$,由 Fisher 计分法可得,$\mu^{(t+1)} = \bar{y}$。与之相对应,Newton-Raphson 算法会给出什么结果?

4.36 利用 Newton-Raphson 算法编写一个计算机程序来最大化二项分布样本的似然函数。在 $n = 10, \hat{\pi} = 0.3$ 时,报告当初始值 $\pi^{(0)}$ 分别为:(a)0.1,(b)0.2,…,(i)0.9 时的前六次迭代结果。讨论初始值对收敛速度的影响。当初始值为 0 或 1 时,会出现什么情况?

4.37 在广义线性模型中,假定对于 $\mu = E(Y)$ 存在 $\text{var}(Y) = v(\mu)$。证明满足 $g'(\mu) = [v(\mu)]^{-1/2}$ 的连结函数 g 在每次迭代过程中具有相同的权数矩阵 $\mathbf{W}^{(t)}$。证明对于泊松分布的随机部分,这个连结函数为 $g(\mu) = 2\sqrt{\mu}$。

4.38 在使用非典型连结的广义线性模型中,证明观察的信息矩阵可能取决于数据本身,并因而不同于期望的信息矩阵。利用 probit 模型,对此加以演示。

5 Logistic 回归

在第 4 章介绍二分数据的广义线性模型时,我们着重强调了 logistic 回归。Logistic 回归是关于分类结果变量的最重要的模型,它在许多不同的领域都有着日益广泛的应用。早期的应用主要集中于生物医学研究,但是在过去的 20 年间,社会科学研究和市场营销中都出现了大量的关于 logistic 回归的应用。

近年来, logistic 回归在商业应用中已经成为一个很普遍的工具。一些*信用评分*(*credit-scoring*)系统利用 logistic 回归对研究对象是否值得信任的概率进行模型分析。例如,可以根据账单的金额、年收入、职业、抵押贷款的负担、过去按时支付账单的百分比,以及申请人信用记录的其他信息等预测变量来估计一个人按时支付账单的概率。依靠发放商品目录进行销售活动的公司可以将对潜在客户销售成功的概率作为其过去购买行为的函数建立模型,从而来决定是否对其发送目录。

对 logistic 回归的另一个日渐兴起的应用领域是遗传学。例如,近期的一篇文章(J. M. Henshall and M. E. Goddard, *Genetics* 151:885-894, 1999)使用 logistic 回归来估计数量特征基因位点效应(quantitative trait loci effects),将子代继承某一种而不是其他等位基因的概率作为不同基因形态表型值的函数进行模型分析。另一篇文章(D. F. Levinson et al., *Amer. J. Hum. Genet.*, 67:652-663, 2000)则通过 logistic 回归对几个研究中心的受感染同胞配对(affected sibling pairs, ASPs)和他们父母的基因型数据进行分析。该模型研究了受感染同胞配对具有同源的等位基因的概率,并检验了不同研究中心之间的异质性。

在本章中,我们将对 logistic 回归进行更详尽的讨论。第 5.1 节介绍如何解释模型中的参数。第 5.2 节介绍有关参数的统计推断。第 5.3 和 5.4 节扩展到多个预测变量的情况,其中有些预测变量可以是分类变量。最后,在第 5.5 节中,我们应用广义线性模型的拟合方法来建立并求解 logistic 回归的似然方程。

5.1 Logistic 回归参数的解释

对于二分结果变量 Y 和解释变量 X,令 $\pi(x)=P(Y=1|X=x)=1-P(Y=0|X=x)$。Logistic 回归模型可表示为:

$$\pi(x)=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}。\tag{5.1}$$

等价地,对数发生比(也称为 Logit)具有以下线性关系:

$$\text{Logit}[\pi(x)]=\log\frac{\pi(x)}{1-\pi(x)}=\alpha+\beta x。\tag{5.2}$$

也即,使二分结果变量的 Logit 连结函数等于线性预测项。

5.1.1 关于 β 的解释:发生比、概率与线性近似

我们应当如何解释公式 5.2 中的 β 呢? 它的符号决定了 $\pi(x)$ 是会随着 x 的增加而上升还是下降,上升或下降的速度随着 $|\beta|$ 的增大而变快。当 $\beta\to 0$ 时,曲线扁平化为一条水平的直线。当 $\beta=0$ 时,Y 独立于 X。对于连续变量 x,在 $\beta>0$ 时,$\pi(x)$ 所对应的曲线具有 logistic 的累积分布函数的形状(回顾第 4.2.5 节)。由于 logistic 密度函数是对称的,$\pi(x)$ 趋向于 1 的速度与它趋向于 0 的速度相同。

对公式 5.2 的两边都求幂表明,发生比是 x 的指数函数。这给出了关于 β 值大小的基本解释。随着 x 每增加一个单位,发生比会增加 e^{β} 倍。换句话说,e^{β} 等于发生比之比,即在 $X=x+1$ 时的发生比除以 $X=x$ 时的发生比。

大多数学者对发生比或者 Logit 并不熟悉,所以解释为关于发生比的 e^{β} 的乘数效应或者关于 Logit 的 β 的可加效应对他们而言没有意义。一种更简单的、近似于斜率的解释利用了线性化的方法(Berkson,1951)。由于 logistic 回归函数(式 5.1)的形状为曲线而不是直线,这意味着,对于 x 的每一单位的变动,在不同的 x 取值处所导致的 $\pi(x)$ 的变动幅度是不一样的。如图 5.1 所示,在 x 的某一特定值与曲线相切的直线描述了曲线在该点的变化速度。利用公式 5.1 计算 $\partial\pi(x)/\partial x$,会得出一个相当复杂的关于参数和 x 的函数,该函数可以简化为 $\beta\pi(x)[1-\pi(x)]$。

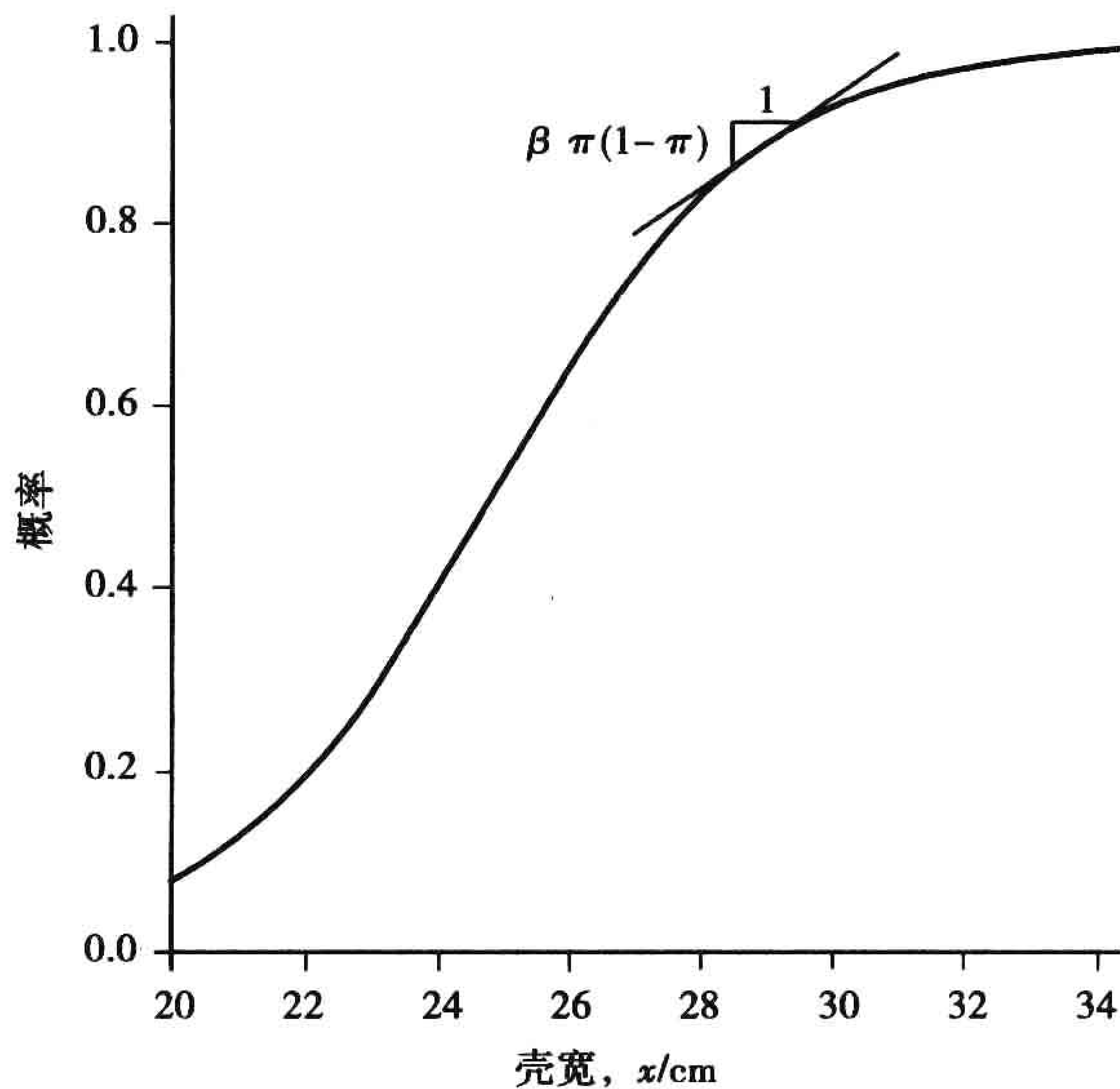


图 5.1 对 logistic 回归曲线的线性近似

例如,在 $\pi(x)=\frac{1}{2}$ 对应的 x 值处,曲线的切线的斜率为 $\beta\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)=\beta/4$;当 $\pi(x)=0.9$ 或 0.1 时,相应的斜率为 0.09β。当 $\pi(x)$ 趋向于 1.0 或 0 时,斜率趋向于 0。最陡的斜率发生在 $\pi(x)=\frac{1}{2}$ 时对应的 x 值处;此时 $x=-\alpha/\beta$(要验证在这一点上 $\pi(x)=\frac{1}{2}$,可以将 $x=-\alpha/\beta$ 代入公式 5.1,或将 $\pi(x)=\frac{1}{2}$ 代入公式 5.2 并对 x 求解)。这个 x 值有时被称为中位效应水平(*median effective level*),并记为 EL_{50}。在毒理学研究中,它被称为 LD_{50}(LD = 致命剂量),即死亡的可能性为 50% 时的剂量。

从这个线性近似可知,当 x 接近于 $\pi(x)=\frac{1}{2}$时,x 每变动 $1/\beta$ 对应着 $\pi(x)$大致变动 $(1/\beta)(\beta/4)=\frac{1}{4}$,也即,$1/\beta$ 近似等于 $\pi(x)=0.25$ 或 0.75(准确地,0.27 和 0.73)对应的 x 值与 $\pi(x)=0.50$ 对应的 x 值之间的距离。然而,只有当 x 的变动较小时,这种线性近似的效果才较好。

对效应的另一种解释办法是计算 x 的某些取值(如 x 的四分位点)所对应的 $\pi(x)$的值。这需要将 x 的四分位点代入公式 5.1 以求出 $\pi(x)$。$\pi(x)$在 x 的中间一半取值上的变动,即从 x 的第一个四分位点到第三个四分位点,描述了 x 的效应。它可以用于与其他预测变量所导致的 $\pi(x)$的相应变动进行比较。

截距参数 α 通常没有什么特别的意义。然而,将预测变量在 0 值进行中心化(centering)(即,用$(x-\bar{x})$取代 x)后,α 就成了在预测变量取均值时所对应的 Logit,因而 $e^{\alpha}/(1+e^{\alpha})=\pi(\bar{x})$(与普通回归一样,在包含二次项或交互项的复杂模型中,中心化有助于降低模型参数估计之间的相关性)。

5.1.2 检查数据

在应用中,对模型的解释需要利用最大似然估计值来替代公式 5.1 中的参数。在拟合模型并对其进行解释之前,我们需要首先检查数据以确定 Logistic 回归模型是否适用。由于 Y 的取值仅为 0 和 1,很难通过直接对 Y 和 x 进行绘图来检查数据。

对样本比例(或 Logit)和 x 进行绘图会有所帮助。令 n_i 表示在 x 取 i 值时的观测值数量,y_i 表示其中结果为“1”的数量,这样,存在 $p_i=y_i/n_i$。第 i 个样本 Logit 等于 $\log[p_i/(1-p_i)]=\log[y_i/(n_i-y_i)]$。当 $y_i=0$ 或 n_i 时,它等于无穷大。这时,可以使用一种特别的调整方法,即在两种类型的结果数量中分别加上一个正的常数。调整后的 Logit 为

$$\log\frac{y_i+\frac{1}{2}}{n_i-y_i+\frac{1}{2}},$$

是相应情况下对真实 Logit 的最小偏差估计(注解 5.2)。关于样本 Logit 的绘图大致应当是线性的。

当 X 是连续变量且所有 $n_i=1$ 时,或当 X 本质上是连续的且所有 n_i 都很小时,上述方法无法给出满意的结果。在计算样本比例和样本 Logit 之前,可以先将相近的 x 值进行合并从而得到分组的数据。不对数据进行合并的一种较好办法是利用修匀技术来揭示数据的趋势。其中一种修匀方法是拟合广义可加模型(第 4.8 节),利用修匀函数来替代广义线性模型中的线性预测项。通过查看拟合图可以发现是否存在对 logistic 回归所预测的 S 型曲线的严重背离。

5.1.3 例子:再论马蹄蟹数据

为了对 logistic 回归进行详细说明,我们重新分析在第 4.3.2 节提到的马蹄蟹数据。这里二分结果变量为雌蟹的周边是否住有同伴:如果至少存在一个同伴,$Y=1$;如果没有,$Y=0$。我们首先使用雌蟹的壳宽作为唯一的预测变量。

图 4.7 显示了一个基于二项分布假设和 Logit 连结的广义可加模型所给出的关于均值的修匀预测。Logistic 回归模型看上去是充分的。在第 4.3.2 节(表 4.4)中用于检查泊松回归模型充分性的分组数据也印证了这一点。针对八个壳宽类别,我们计算出每个

类别中拥有同伴的马蹄蟹所占的样本比例以及该类别的平均壳宽。图 5.2 显示了拥有同伴的雌蟹的样本比例与相应的壳宽均值所对应的八个点。图中八个样本比例与广义可加模型的修匀曲线都显示出大致上升的趋势,因此我们将壳宽作为线性预测变量拟合 logistic 回归模型。

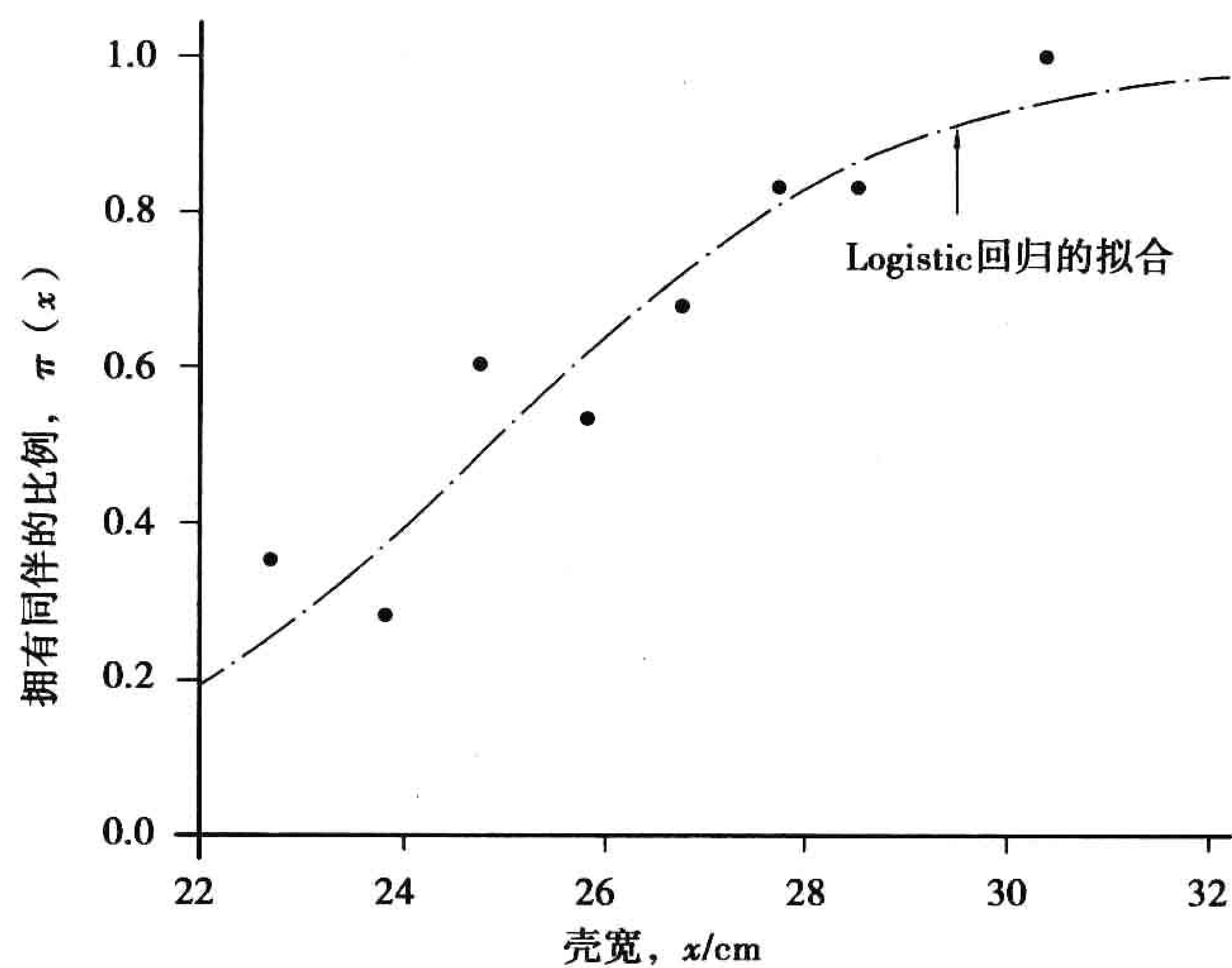


图 5.2　按壳宽划分的拥有同伴的雌蟹比例的观察值和拟合值

关于最大似然的拟合,我们将留在第 5.5 节再具体讨论。表 5.1 所示为软件(如,关于 SAS 参见表 A.8)给出的结果。在表 4.3 的未分组数据中,令 $\pi(x)$ 表示壳宽为 x 的雌性马蹄蟹拥有同伴的概率。最大似然拟合结果为

$$\hat{\pi}(x) = \frac{\exp(-12.351 + 0.497x)}{1 + \exp(-12.351 + 0.497x)}。$$

将整个样本的均值 $x = 26.3$ 厘米代入拟合方程,得到 $\hat{\pi}(x) = 0.674$。当 $x = -\hat{\alpha}/\hat{\beta} = 12.351/0.497 = 24.8$ 时,估计的概率等于$\frac{1}{2}$。图 5.2 显示了不同壳宽所对应的 $\hat{\pi}(x)$。

表 5.1　关于马蹄蟹数据的 Logistic 回归模型的输出结果

Criteria For Assessing Goodness of Fit

Criterion	DF	Value
Deviance	171	194.452 7
Pearson Chi-Square	171	165.143 4
Log Likelihood		−97.226 3

Parameter	Estimate	Std Error	Likelihood-Ratio 95% Conf Limits		Wald Chi-Sq	P > ChiSq
Intercept	−12.350 8	2.628 7	−17.809 7	−7.457 3	22.07	<.000 1
Width	0.497 2	0.101 7	0.308 4	0.709 0	23.89	<.000 1

译者注——Width:壳宽。

壳宽每增加 1 厘米,估计的拥有同伴的发生比增加 $\exp(\hat{\beta}) = \exp(0.497) = 1.64$ 倍,也即,上升 64%。为了更通俗地表述这一效应,我们可以报告拥有同伴的概率的变动速度。在壳宽为均值时,$\hat{\pi}(x) = 0.674$,并且壳宽每增加 1 厘米,$\hat{\pi}(x)$ 大约增加 $\hat{\beta}[\hat{\pi}(x)(1-\hat{\pi}(x))] = 0.497(0.674)(0.326) = 0.11$。或者,我们可以报告 x 的四分位点所对应

的 $\hat{\pi}(x)$。壳宽的第一个四分位点、中位数和第三个四分位点分别为 24.9,26.1 和 27.7;在这些值上 $\hat{\pi}(x)$ 分别等于 0.51,0.65 和 0.81,即在 x 的取值位于中间一半的样本中 $\hat{\pi}(x)$ 增加了0.30。

上述最后一种解释可用于比较具有不同度量单位的预测变量的效应。例如,利用蟹的体重作为预测变量,拟合结果为 $\text{Logit}[\hat{\pi}(x)] = -3.695 + 1.815x$。由于体重增加 1 千克与壳宽增加 1 厘米之间没有可比性,所以 x = 壳宽时的 $\hat{\beta} = 0.497$ 与 x = 体重时的 $\hat{\beta} = 1.815$ 是不能直接进行对比的。体重的四分位点分别为 2.00,2.35 和 2.85,对应的 $\hat{\pi}(x)$ 为 0.48,0.64 和 0.81,在体重的取值位于中间一半的样本中 $\hat{\pi}(x)$ 增长了 0.33。这个效应与壳宽的效应大致相当。

5.1.4 回顾性研究中的 Logistic 回归

Logistic 回归的另一个特性关系到当解释变量 X(而不是结果变量 Y)是随机的情形。这种情况发生在回顾性抽样设计,比如个案-控制的生物医药研究中(第 2.1.6 节)。由具有 $Y=1$(个案)和 $Y=0$(控制)的研究对象组成样本,对 X 的值进行观测。如果在个案和控制之间 X 值的分布是有差别的,就表明二者之间存在关联。在回顾性研究中,我们可以估计发生比之比(第 2.2.4 节)。Logistic 回归模型中的效应与发生比之比有关,因此,可以通过拟合该模型来估计个案-控制研究中的效应。

以下解释了这样做的依据。令 Z 表示一个对象是否被选中(1 = 是,0 = 否),则有 $\rho_1 = P(Z=1 \mid y=1)$ 表示选中一个个案的概率,$\rho_0 = P(Z=1 \mid y=0)$ 表示选中一个控制案例的概率。尽管并未对在给定 $X=x$ 的情况下 Y 的条件分布进行抽样,但是我们需要关于 $P(Y=1 \mid z=1,x)$ 的模型,假定 $P(Y=1 \mid x)$ 服从 logistic 模型。按照贝叶斯定理,

$$P = (Y = 1 \mid z = 1,x) = \frac{P(Z = 1 \mid y = 1,x)P(Y = 1 \mid x)}{\sum_{j=0}^{1} P(Z = 1 \mid y = j,x)P(Y = j \mid x)}。\qquad (5.3)$$

现在,假定对于 $y=0$ 和 1,$P(Z=1 \mid y,x) = P(Z=1 \mid y)$,也即,对于每一个 y,抽样概率不取决于 x。例如,x 常常是指某种形式的历险因素,比如某人是否吸烟。这时,对于个案和控制来说,吸烟者和非吸烟者被选中的概率是一样的。在这个假定下,将 ρ_1 和 ρ_0 代入公式 5.3,并在分子和分母中同时除以 $P(Y=0 \mid x)$,公式 5.3 便简化为

$$P = (Y = 1 \mid z = 1,x) = \frac{\rho_1 \exp(\alpha + \beta x)}{\rho_0 + \rho_1 \exp(\alpha + \beta x)}。$$

在上式分子和分母中再都除以 ρ_0,并代入 $\rho_1/\rho_0 = \exp[\log(\rho_1/\rho_0)]$,可得:

$$\text{Logit}[P(Y = 1 \mid z = 1,x)] = \alpha^* + \beta x,$$

其中 $\alpha^* = \alpha + \log(\rho_1/\rho_0)$。

因此,logistic 回归模型中的效应参数 β 与在模型 $P(Y=1 \mid x)$ 中是相同的。如果个案的抽样率是控制组的 10 倍,logistic 回归模型所估计的截距比通过前瞻性研究估计的截距大 $\log(10) = 2.3$ 倍。有关评述,参见:Anderson(1972)、Breslow and Day(1980,p. 203)、Breslow and Powers(1978)、Carroll et al.(1995)、Farewell(1979)、Mantel(1973)、Prentice(1976a)、Prentice and Pyke(1979)。

在个案-控制研究中,我们无法利用其他二分结果变量模型来估计 β。与发生比之比不同,给定 Y 时 X 的条件分布的效应不同于给定 X 时 Y 的条件分布的效应。这是 Logit 连结的重要优点,也是在生物医学研究中 Logit 模型占据压倒性优势的一个主要原因。

许多个案-控制研究使用了配对的方法,即对每个个案匹配一个或多个控制对象。控

制对象和个案对象在主要特征(比如年龄)上是一样的。模型构建以及随后的分析应当将配对的情况考虑进来。有关配对的个案-控制研究的 logistic 回归,我们将留在第10.2.5节讨论。

不论是何种抽样设计,logistic 回归可能可以很好地描述所研究的关系,也可能不行,但在以下这种特定情况下,它一定成立。给定 $Y=i$,假如 X 服从 $N(\mu_i,\sigma^2)$ 分布,其中 $i=0,1$。那么,按照贝叶斯定理,$P(Y=1 \mid X=x)$ 等于公式 5.1,其中 $\beta=(\mu_1-\mu_0)/\sigma^2$ (Cornfield,1962)。当一个总体是由两种类型的对象混合而成时,其中一种为 $Y=1$,X 近似于正态分布,另一种为 $Y=0$,X 也近似于正态分布,并且二者方差相似。这时,公式5.1的 logistic 回归函数能很好地近似曲线 $\pi(x)$。如果分布是正态的,但是方差不同,可以使用包括二次项的 logistic 回归模型(Anderson,1975)。在该情况下,存在着非单调的关系,$\pi(x)$ 会先上升再下降,或者相反(习题 5.33)。

5.2 Logistic 回归的统计推断

按照沃尔德(Wald,1943)关于最大似然估计值的渐近结果,logistic 回归模型中的参数估计值服从大样本正态分布。因而,可以通过沃尔德、似然比和计分这三种方法(第1.3.3节)进行统计推断。

5.2.1 统计推断的类型

对于只有一个预测变量的模型,

$$\text{Logit}[\pi(x)] = \alpha + \beta x,$$

显著性检验的焦点是 $H_0:\beta=0$,即独立性假设。沃尔德检验运用 $\hat{\beta}$ 对应的对数似然函数,检验统计量为 $z=\hat{\beta}/\text{SE}$ 或 z 的平方;在 H_0 下,z^2 渐近于 χ_1^2。似然比检验运用 $\hat{\beta}$ 与 $\beta=0$ 时的最大对数似然值之差的两倍,它也渐近服从 χ_1^2 零分布。计分检验运用对数似然函数在 $\beta=0$ 处的导数(即,计分函数)。该检验统计量将 β 的充分统计量与其零期望值相比较,并进行适当的标准化[$N(0,1)$ 或 χ_1^2]。关于 $H_0:\beta=0$ 的计分检验的详细内容,我们将在第 5.3.5 节介绍。

在大样本的情况下,上述三种检验往往会给出相似的结果。比较而言,由于似然比检验综合考虑了在 H_0 和在 $\hat{\beta}$ 时的对数似然函数,利用了更多的信息,因而似然比检验要优于沃尔德检验。当 $|\beta|$ 相对较大时,沃尔德检验不如似然比检验那么有效,甚至有可能出现异常结果(参见:Hauck and Donner(1977),以及习题 5.38)。

与检验本身相比,置信区间给出了更多的信息。关于 β 的区间可以通过对 $H_0:\beta=\beta_0$ 的检验过程的逆运算来得到。相应区间就是满足卡方检验统计量不大于 $\chi_1^2(\alpha)=z_{\alpha/2}^2$ 的一组 β_0。就沃尔德方法而言,这意味着 $[(\hat{\beta}-\beta_0)/\text{SE}]^2 \leqslant z_{\alpha/2}^2$,对应的置信区间为 $\hat{\beta} \pm z_{\alpha/2}$(SE)。

为概括所研究的关系,其他的数据特征可能提供比 β 更重要的信息,比如在不同的 x 取值处所对应的 $\pi(x)$ 的值。对于给定的 $x=x_0$,$\text{Logit}[\hat{\pi}(x_0)]=\hat{\alpha}+\hat{\beta}x_0$ 的大样本标准误(SE)可由下式的平方根来估计:

$$\text{var}(\hat{\alpha}+\hat{\beta}x_0) = \text{var}(\hat{\alpha}) + x_0^2\text{var}(\hat{\beta}) + 2x_0\text{cov}(\hat{\alpha},\hat{\beta})。$$

关于 $\text{Logit}[\pi(x_0)]$ 的 95% 的置信区间为 $(\hat{\alpha}+\hat{\beta}x_0) \pm 1.96\text{SE}$。将两个端点分别代入逆转

换 $\pi(x_0)=\exp(\text{Logit})/[1+\exp(\text{Logit})]$,便可得出关于 $\pi(x_0)$ 的相应区间。

上述每一种统计推断方法都可以给出小样本的置信区间和检验。对此,我们将留在第 6.7 节讨论。

5.2.2 以马蹄蟹数据为例进行统计推断

为了演示 logistic 回归的统计推断,我们建立以壳宽为预测变量来预测马蹄蟹拥有同伴的概率的模型。表 5.1 给出了模型的拟合结果和标准误。统计量 $z=\hat{\beta}/SE=0.497/0.102=4.9$ 对存在正的壳宽效应提供了强有力的证据($P<0.0001$)。与之相等价的沃尔德卡方统计量为 $z^2=23.9$,相应自由度为 df = 1。在 $H_0:\beta=0$ 下的最大对数似然值等于 -112.88,完全模型的最大对数似然值为 -97.23。似然比统计量等于 $-2(-112.88-97.23)=31.3$,相应自由度为 df = 1。由此可见,似然比检验给出了比沃尔德检验更强的证据。

关于 β 的 95% 的沃尔德置信区间为 $0.497\pm1.96(0.102)$,即(0.298,0.697)。表 5.1给出的基于剖面似然函数的似然比置信区间为(0.308,0.709)。壳宽每增加 1 厘米,对发生比的效应的置信区间为$(e^{0.308},e^{0.709})=(1.36,2.03)$。我们推论说,壳宽每增加 1 厘米至少会使拥有同伴的发生比上升 30%,最多会使其翻倍。

大多数关于 logistic 回归的软件也报告 $\pi(x)$ 的估计及其置信区间(例如,在 SAS 的 PROC GENMOD 中选择 OBSTATS 选项)。以壳宽接近于均值(即 $x=26.5$)的情况为例。模型估计的 Logit 为 $-12.351+0.497(26.5)=0.825$,且 $\hat{\pi}(x)=0.695$。软件同时输出

$$\widehat{\text{var}}(\hat{\alpha})=6.910,\quad \widehat{\text{var}}(\hat{\beta})=0.01035,\quad \widehat{\text{cov}}(\hat{\alpha},\hat{\beta})=-0.2668,$$

由此可得:

$$\widehat{\text{var}}\{\text{Logit}[\hat{\pi}(x)]\}=6.910+x^2(0.01035)+2x(-0.2668)。$$

在 $x=26.5$ 时,上式等于 0.038,因此,关于 Logit[$\pi(26.5)$] 的 95% 的置信区间等于 $0.825\pm(1.96)\sqrt{0.038}$,即(0.44,1.21)。这可以转换为拥有同伴的概率的置信区间(0.61,0.77)(如 $\exp(0.44)/[1+\exp(0.44)]=0.61$)(还有一种方法,利用 $x^*=x-26.5$作为预测变量来拟合模型,这时 $\hat{\alpha}$ 和它的标准误就是估计的 Logit 及其标准误)。图 5.3 显示了将 $\pi(x)$ 视为 x 的函数对应的预测方程的置信边界。Hauck(Hauck,1983)介绍了另一种边界,其中置信系数可同时应用于预测变量的所有可能取值。

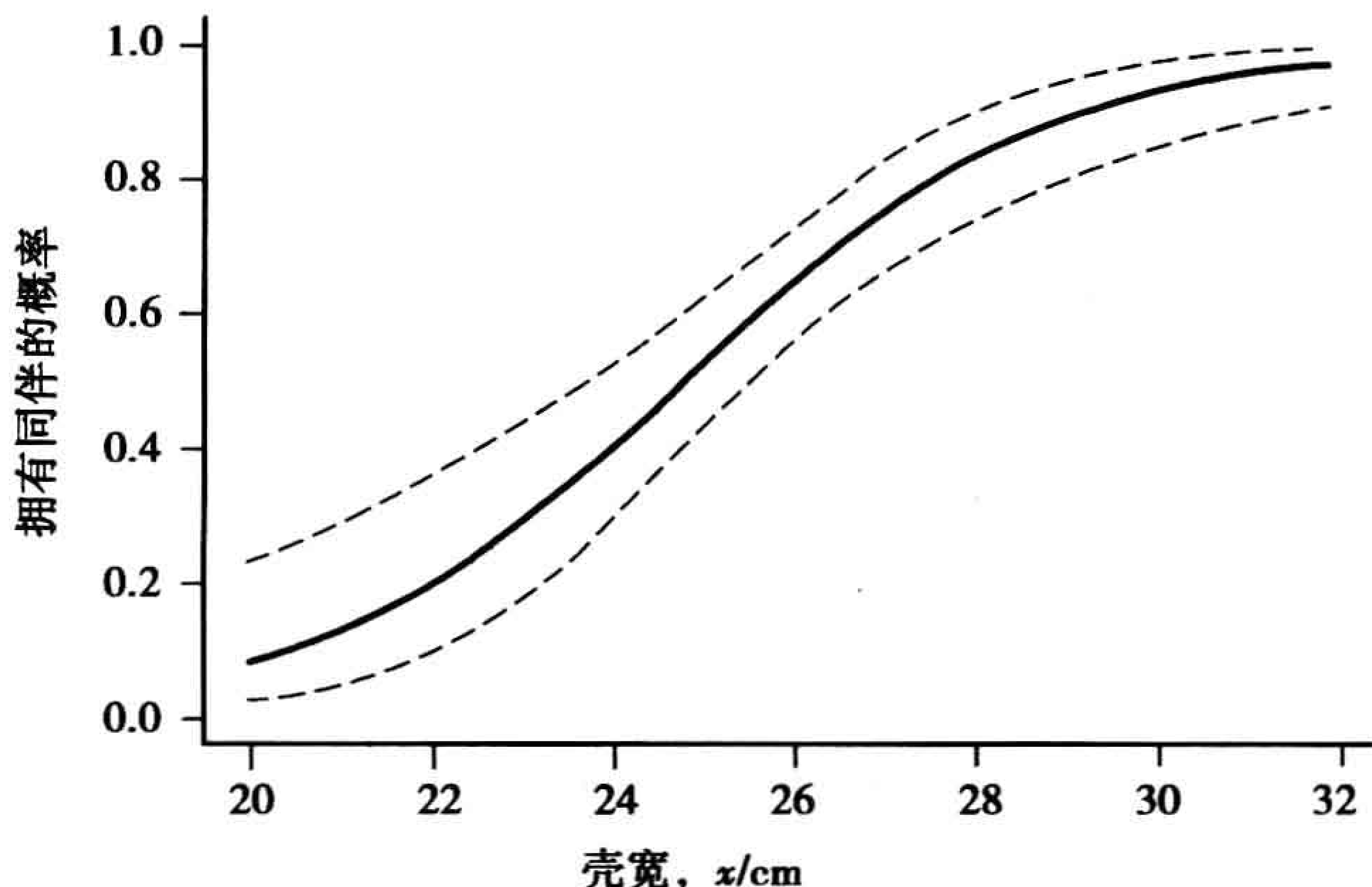


图 5.3 利用壳宽预测拥有同伴的概率的预测方程和 95% 置信边界

大家也可以忽略模型拟合而仅使用样本比例(即饱和模型)来估计这些概率。样本

中有6只雌蟹的 $x=26.5$,其中4只拥有同伴。在 $x=26.5$ 时的样本比例估计值为 $\hat{\pi}=4/6=0.67$,这与模型估计值相似。单独根据这6个观测值得出的95%的计分置信区间(第1.4.2节)等于(0.30,0.90)。

当logistic回归模型成立时,关于概率的模型估计值明显优于样本比例。这时,模型仅需要估计两个参数,而饱和模型需要对每个 x 的不同取值都逐一估计单独的参数。例如,在 $x=26.5$ 时,软件输出结果中模型估计值0.695的标准误为0.04,而只根据6个观测值所计算的样本比例的标准误为 $\sqrt{\hat{\pi}(1-\hat{\pi})/n}=\sqrt{(0.67)(0.33)/6}=0.19$。使用模型得到的95%的置信区间是(0.61,0.77),相比之下,使用样本比例的置信区间为(0.30,0.90)。与样本比例中仅仅使用6个观测值相反,模型利用了所有173个观测值的信息来估计两个模型参数,因而其估计结果要精确得多。

然而,现实情况往往要更复杂一些。在实际应用中,模型并不就是 $\pi(x)$ 与 x 之间的*精确*关系。如果模型对真实概率的近似能令人满意,其估计值仍然比样本比例更接近于真实值。模型对样本数据进行了修匀,从而在一定程度上降低了观察到的变动性。除非每个样本比例都是基于极大的样本规模,模型估计值一般都优于样本比例。在第6.4.5节中,我们将进一步讨论模型分析的这一优点。

5.2.3 检查拟合优度:未分组数据和分组数据

在现实中,并不能保证某一logistic回归模型就能对数据进行很好的拟合。对于任何类型的二分数据,检查模型拟合不足的方法之一是利用似然比检验将该模型与比之更复杂的模型进行对比。更复杂的模型可能包括非线性的效应,比如二次项。具有多个预测变量的模型可以考虑交互效应。如果更复杂的模型未能对数据拟合得更好,这就对所选模型的合理性提供了一定的支持。

检查模型拟合不足的其他方法可以是寻找模型出现问题的*任何*方式。当解释变量都是分类变量时,这种情况最为简单(如我们将在第5.4.3节所讨论的那样)。在每个 x 的取值水平上,将两种结果的估计概率乘以在该取值处的对象数量可以得出关于 $y=0$ 和 $y=1$ 的期望频数的估计值。这些是模型的*拟合值*(*fitted values*)。模型检验可以利用皮尔逊 X^2 或似然比 G^2 统计量来比较观察计数和拟合值之间的差别。当 x 的类别数量一定时,随着拟合计数的增加,X^2 和 G^2 的极限服从卡方零分布。分布的自由度,也称为模型的*残差*(*residual*)自由度,等于饱和模型的参数个数(即,x 的类别数)减去该模型的参数个数。

在这里,我们将模型拟合的整体检验局限于分类预测变量的情况,其原因在于第4.5.3节提及的关于二项分布模型中分组数据与未分组数据的区分。在这两种情况下,饱和模型是有区别的。只有在饱和模型的参数个数固定,也即预测变量的类别数量为给定的情况下,当 $n\to\infty$ 时偏离度才渐近服从卡方分布。

5.2.4 马蹄蟹数据的模型拟合优度

我们通过以 $x=$ 壳宽来预测雌蟹拥有同伴的概率的模型,来演示有关模型拟合优度的分析。检查拟合优度的一种方法是将其与一个更复杂的模型进行比较,如一个包括二次项的模型。通过将壳宽减去它的均值26.3对其在0值进行中心化,这样,模型的拟合结果为

$$\text{Logit}[\hat{\pi}(x)]=0.618+0.533x+0.040x^2。$$

二次项估计的标准误 SE = 0.046。检验 x^2 的真实参数为 0 的似然比统计量等于 0.83（df = 1），没有什么证据支持增加这一项。

接下来，我们考虑模型总体的拟合优度。在 173 只蟹中壳宽共有 66 个不同取值，在多数壳宽取值处只对应极少的观测值。可以将这个数据视为一个 66 × 2 的列联表。每一行的两个单元格分别给出在这个壳宽水平上拥有同伴和没有同伴的蟹的数量。当 x 的不同取值水平的数量固定，且在每一取值处的观测值数量增加时，可以应用有关 X^2 和 G^2 的卡方理论。尽管我们利用壳宽的不同取值对数据进行了分组而不是应用原始的 173 个二分观测值，这里还是在两个方面不满足上述理论：第一，大多数的拟合计数都非常小；第二，当收集更多的数据时，壳宽可以取其他的值，因而列联表会包括更多的而不是固定数目的单元格。因此，对于包括连续或接近于连续的预测变量的 logistic 回归模型，其 X^2 和 G^2 不近似于卡方分布（这种情况下，正态性近似可能更适当，但是目前还没有一种方法受到特别的关注；相应文献，参见第 9.8.6 节）。

利用 X^2 和 G^2 可以比较分组数据中的观察值和拟合值。表 5.2 使用了与表 4.4 相同的分组，给出了一个 8 × 2 表格。对每一个壳宽的类别，结果为"有"的拟合值等于壳宽落在该类别的所有蟹的估计概率 $\hat{\pi}(x)$ 的加总；结果为"没有"的拟合值等于对这些蟹的 $1-\hat{\pi}(x)$ 的加总。这样得到的拟合值会大得多，因而，X^2 和 G^2 更为有效，尽管由于 $\pi(x)$ 在每一类别内并不恒定，这就使得卡方理论并不完全适用。相应结果分别为 $X^2 = 5.3$ 以及 $G^2 = 6.2$。对应于每个壳宽类别，表 5.2 包括了八个二项分布样本；由于模型具有两个参数，因而拟合检验的自由度 df = 8 − 2 = 6。X^2 和 G^2 都没有显示存在拟合不足的问题（$P > 0.4$），因此，最初的对未分组数据的拟合结果得到了进一步支持。

表 5.2　拟合 Logistic 回归模型的马蹄蟹分组数据的观察值和拟合值

壳宽/cm	"有"的数量	"没有"的数量	"有"的拟合值	"没有"的拟合值
<23.25	5	9	3.64	10.36
23.25 ~ 24.25	4	10	5.31	8.69
24.25 ~ 25.25	17	11	13.78	14.22
25.25 ~ 26.25	21	18	24.23	14.77
26.25 ~ 27.25	15	7	15.94	6.06
27.25 ~ 28.25	20	4	19.38	4.62
28.25 ~ 29.25	15	3	15.65	2.35
>29.25	14	0	13.08	0.92

5.2.5　通过对未分组数据进行分组来检查拟合优度

如上所述，对于未分组数据，以及存在连续或近似连续的预测变量的情况，X^2 和 G^2 的极限不服从卡方分布。它们仍然可以用于比较模型，如上一节中对二次项的检查以及我们将在第 5.4.3 节和第 9.8.5 节讨论的情况。另外，如上文所指出的，X^2 和 G^2 可以近似地应用于对 x 的取值空间进行分割所产生的分组数据的观察值和拟合值。然而，随着解释变量数量的增加，同时对每个变量的值进行分组会导致列联表的单元格数量急剧膨大，这样，其中许多单元格的计数都会很小。

不论有多少预测变量，都可以按照由原始的未分组数据估计的成功概率来分割观察

值和拟合值。一种通用的办法是使分割所形成的组别规模大致相等。以10组为例，第一组的观察计数和相应的拟合计数是具有最高的估计概率的 $n/10$ 个观测值，下一组包括估计概率为第2个十分位点的 $n/10$ 个观测值，依此类推。对结果变量的每一种结果，各组都包括一个观察计数及其拟合值。对某种结果的拟合值等于该组中所有观测值出现该结果的估计概率的加总。

基于上述原理，Hosmer 和 Lemeshow(1980)提出了一个通过这种分割来比较观察计数和拟合计数的皮尔逊统计量。令 y_{ij} 表示分割在第 i 组的第 j 个观测值的二分结果，其中 $i=1,\cdots,g,j=1,\cdots n_i$。令 $\hat{\pi}_{ij}$ 表示利用未分组数据拟合的模型中相应的概率。他们的统计量可表示为

$$\sum_{i=1}^{g}\frac{(\sum_j y_{ij}-\sum_j\hat{\pi}_{ij})^2}{(\sum_j\hat{\pi}_{ij})[1-(\sum_j\hat{\pi}_{ij})/n_i]}。$$

当许多观测值具有相同的估计概率时，对这些观测值进行分组具有一定的随意性，这时，不同的软件可能给出不同的值。该统计量的极限不服从卡方分布，这是因为每一组中的观测值不是完全等同的试验，它们也不具有相同的成功概率。但是，Hosmer 和 Lemeshow 指出，当协变量的不同取值组合的数量等于样本规模时，零分布近似服从自由度 $df=g-2$ 的卡方分布。

将壳宽视为连续的预测变量，对马蹄蟹数据拟合 logistic 回归模型。在 $g=10$ 组的情况下，Hosmer-Lemeshow 统计量等于3.5，自由度为 $df=8$。这也表明模型拟合结果可以接受。

遗憾的是，与其他对模型总体拟合进行检验的统计量一样，Hosmer-Lemeshow 统计量对于发现某些特定类型的拟合不足统计效能较差(Hosmer et al.，1997)。在任何情况下，模型总体拟合检验统计量的值很大只是表明存在*某些*拟合不足，却并不告诉我们其来源。就科学性而言，将现有模型与更复杂的模型进行比较的方法更有意义，因为它试图寻求某一种特定类型的拟合不足。无论使用哪种方法，当存在拟合不足时，可以通过模型诊断工具描述单个观测值对模型拟合结果的影响并给出拟合不充分的原因。我们将在第6.2.1节对此进行讨论。

5.3 包括分类预测变量的 Logit 模型

与普通回归一样，logistic 回归也可以扩展到包括定性解释变量(常称为*因子*(*factors*))的情况。在本节中，我们利用虚拟变量来完成这一扩展。

5.3.1 对因子的 ANOVA 表述

为简单起见，我们首先考虑单个因子 X 的情况，其具有 I 个类别。在 $I\times2$ 表格的第 i 行，y_i 是在 n_i 次试验中出现第一列结果(成功)的次数。我们将 y_i 看作参数为 π_i 的二项分布变量。

仅包括一个因子的 Logit 模型可表示为

$$\log\frac{\pi_i}{1-\pi_i}=\alpha+\beta_i。\tag{5.4}$$

β_i 越大，π_i 的值也就越大。公式5.4的右边与单元格均值的单方差分析(one-way ANOVA)的模型公式相类似。与方差分析相同，因子所具有的参数 $\{\beta_i\}$ 的数量等于它的

类别数，但其中一个参数是冗余的。当存在 I 个类别时，X 具有 $I-1$ 个非冗余的参数。冗余参数可以设定为 0，如 $\beta_I=0$。如果这些值不满足这一限定条件，我们可以对其重新编码。例如，设定 $\tilde{\beta}_i=\beta_i-\beta_I$ 以及 $\tilde{\alpha}=\alpha+\beta_I$，这样就满足 $\tilde{\beta}_I=0$。这时，

$$\text{Logit}(\pi_i)=\alpha+\beta_i=(\tilde{\alpha}-\beta_I)+(\tilde{\beta}_i+\beta_I)=\tilde{\alpha}+\tilde{\beta}_i,$$

其中新定义的参数满足限定条件。当 $\beta_I=0$ 时，α 等于第 I 行的 Logit，并且 β_i 是第 i 行和第 I 行的 Logit 之差。因此，β_i 等于这两行的对数发生比之比。

对于*任意的* $\{\pi_i>0\}$，都存在 $\{\beta_i\}$ 使得模型（式 5.4）成立。这个模型具有与二项分布个数一样多的参数，因而是*饱和*模型。当因子不存在效应时，$\beta_1=\beta_2=\cdots=\beta_I$。这等价于 $\pi_1=\cdots=\pi_I$，这时，只包括一个截距项的模型表明 X 和 Y 在统计上独立。

5.3.2 Logit 模型中的虚拟变量

通过*虚拟变量*（*dummy variables*），可以对模型（式 5.4）进行等价表述。对于第 i 行的观测值，令 $x_i=1$，否则 $x_i=0$，其中 $i=1,\cdots,I-1$。这样，模型可表示为

$$\text{Logit}(\pi_i)=\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_{I-1}x_{I-1}。$$

考虑到冗余参数的问题，我们没有包括关于第 I 个类别的虚拟变量。模型（式 5.4）中的限定条件 $\beta_I=0$ 对应着这里所使用的虚拟变量。对于虚拟变量而言，选择不包括哪一个类别是随意的。一些软件设定 $\beta_1=0$，这就对应于仅包括关于第 2 个至第 I 个类别的虚拟变量而不包括第 1 个类别的模型。

另一种设置限制条件的方法是令 $\sum_i\beta_i=0$。假定 X 具有 $I=2$ 个类别，这时存在 $\beta_1=-\beta_2$。这对应于对虚拟变量的如下*效应编码*（*effect coding*），第1类为 $x=1$，第2类为 $x=-1$。

无论使用哪种编码规则，实质的结论都是一样的。对于模型（式 5.4），无论怎样限定 $\{\beta_i\}$，$\{\hat{\alpha}+\hat{\beta}_i\}$ 以及 $\{\hat{\pi}_i\}$ 都是相同的。X 的两个类别（a,b）所对应的差 $\hat{\beta}_a-\hat{\beta}_b$ 也完全相同，它代表了估计的对数发生比之比。因此，$\exp(\hat{\beta}_a-\hat{\beta}_b)$ 表示在 X 的第 a 类别所估计的成功的发生比除以在 X 的第 b 类别所估计的成功的发生比。对一个模型重新设定参数（reparameterizing）或许会使参数的估计值发生变化，但并不会改变模型的拟合结果以及所关注效应的大小。

由于不同的限定条件会导致不同的值，单个类别的 β_i 或 $\hat{\beta}_i$ 的值是没有意义的。例如，对于一个二分预测变量，以 $\beta_2=0$ 为参照组的虚拟变量，对数发生比之比等于 $\beta_1-\beta_2=\beta_1$；相反，对于效应编码为 ±1 因而 $\beta_1+\beta_2=0$ 的虚拟变量，对数发生比之比等于 $\beta_1-\beta_2=\beta_1-(-\beta_1)=2\beta_1$。一个参数或其估计值只有在与另一个类别的参数进行比较时才有意义。

5.3.3 例子：再论饮酒与婴儿畸形

现在，我们回到表 3.7 关于孕妇饮酒与婴儿先天畸形的研究，表 5.3 再次给出了相应数据。在模型（式 5.4）中，我们将畸形作为结果变量，饮酒作为解释因子。无论如何对 $\{\beta_i\}$ 进行限定，$\{\hat{\alpha}+\hat{\beta}_i\}$ 都是样本 Logit，如表 5.3 所示。例如，

$$\text{Logit}(\hat{\pi}_1)=\hat{\alpha}+\hat{\beta}_1=\log(48/17\ 066)=-5.87。$$

当设定 $\beta_5=0$ 时，$\hat{\alpha}=-3.61$，$\hat{\beta}_1=-2.26$。当设定 $\beta_1=0$ 时，$\hat{\alpha}=-5.87$。如表 5.3 所示，除第一个和第二个类别之间存在微小的不一致外，Logit 以及发生畸形的样本比例都随着

饮酒量的增加而增加。

设定所有 $\beta_i=0$ 的简单模型对应的是独立性的情况。对此，$\hat{\alpha}$ 等于畸形的所有样本比例的 Logit，即 log(93/32 481) = −5.86。对“H_0：独立性”进行检验(df = 4)，皮尔逊统计量(式 3.10)为 $X^2=12.1(P=0.02)$，似然比统计量(式 3.11)为 $G^2=6.2(P=0.19)$。这两个检验给出了相反的结论。表 5.3 中有的计数非常小，有的中等，还有一些非常大。这样，即便 $n=32\ 574$，X^2 或 G^2 的样本零分布可能也不近似于卡方分布。利用 X^2 和 G^2 的精确条件分布求得的 P 值分别为 0.03 和 0.13，二者更接近一些，但是给出的结论仍然不同。无论哪种情况，这些统计量忽略了饮酒量的排序特征。样本显示较大的饮酒量更倾向于导致畸形。头两个百分点很相似，接下去的两个也相近，然而，最后三个百分点都会因是否删除一个畸形案例的观测值而发生明显变化。

表 5.3　关于表 3.7 中畸形的 Logit 和样本比例

饮酒量	出现畸形	未出现畸形	Logit	畸形的比例	
				观察值	拟合值
0	48	17 066	−5.87	0.002 8	0.002 6
<1	38	14 464	−5.94	0.002 6	0.003 0
1~2	5	788	−5.06	0.006 3	0.004 1
3~5	1	126	−4.84	0.007 9	0.009 1
≥6	1	37	−3.61	0.026 3	0.023 1

5.3.4　$I\times2$ 表格的线性 Logit 模型

模型(式 5.4)将解释因子视为定类变量，相应类别间的排序没有实际意义。对于定序因子，存在比模型(式 5.4)更为简约，但比独立性模型更复杂的模型。例如，令赋值 $\{x_1,x_2,\cdots x_I\}$ 描述 X 的不同类别之间的距离，如果预期 X 对 Y 的效应是单调的，那么通常的做法是拟合*线性 Logit 模型*(*linear Logit model*)：

$$\text{Logit}(\pi_i)=\alpha+\beta x_i。\tag{5.5}$$

独立性模型是上述模型中 $\beta=0$ 时的特例。

在表 5.3 中样本 Logit 之间近乎单调的增加趋势表明，线性 Logit 模型(式 5.5)可能比独立性模型拟合得更好。按照饮酒量的测量方式，它将一个自然的连续变量进行了分组。通过赋值 $\{x_1=0,x_2=0.5,x_3=1.5,x_4=4.0,x_5=7.0\}$，其中最后一个赋值具有一定的随意性，表 5.4 给出了相应模型的拟合结果。模型估计结果显示，日饮酒量每增加一个单位对畸形的发生比的可积效应为 exp(0.317) = 1.37。表 5.3 给出了畸形发生比例的观察值和拟合值。模型似乎拟合得很好，如对比观察计数与拟合计数的统计量为 $G^2=1.95$ 和 $X^2=2.05$，自由度为 df = 3。

5.3.5　Cochran-Armitage 趋势检验

Armitage(1955)和 Cochran(1954)是最早强调在列联表分析中利用类别间排序信息的重要性的学者之一。对于行变量为定序变量的 $I\times2$ 表格以及 I 个服从 $\text{bin}(n_i,\pi_i)$ 的独立变量 $\{y_i\}$，Armitage 和 Cochran 提出了通过分割皮尔逊统计量来进行独立性检验的趋势统计量。他们使用了线性概率模型，

$$\pi_i = \alpha + \beta x_i, \tag{5.6}$$

表 5.4 对婴儿畸形数据拟合 Logistic 回归模型的输出结果

Criteria For Assessing Goodness of Fit

Criterion	DF	Value
Deviance	3	1.948 7
Pearson Chi-Square	3	2.052 3
Log Likelihood		-635.596 8

Parameter	Estimate	Std Error	Likelihood-Ratio 95% Conf Limits		Wald Chi-Sq	Pr > ChiSq
Intercept	-5.960 5	0.115 4	-6.193 0	-5.739 7	2 666.41	<.000 1
alcohol	0.316 6	0.125 4	0.018 7	0.523 6	6.37	0.011 6

该模型可以通过普通最小二乘法来拟合。在这个模型中,独立性的零假设是 $H_0:\beta = 0$。令 $\bar{x} = \sum_i n_i x_i / n, p_i = y_i / n_i$,并令 $p = (\sum_i y_i)/n$ 表示总的成功比例。预测方程为

$$\hat{\pi}_i = p + b(x_i - \bar{x}),$$

其中

$$b = \frac{\sum_i n_i (p_i - p)(x_i - \bar{x})}{\sum_i n_i (x_i - \bar{x})^2}。$$

记检验独立性的皮尔逊统计量为 $X^2(I)$。对于行变量为定序变量的 $I \times 2$ 表格,满足

$$X^2(I) = \frac{1}{p(1-p)} \sum_i n_i (p_i - p)^2 = z^2 + X^2(L),$$

其中

$$X^2(L) = \frac{1}{p(1-p)} \sum_i n_i (p_i - \hat{\pi}_i)^2,$$

$$z^2 = \frac{b^2}{p(1-p)} \sum_i n_i (x_i - \bar{x})^2 = \left[\frac{\sum_i (x_i - \bar{x}) y_i}{\sqrt{p(1-p) \sum_i n_i (x_i - \bar{x})^2}} \right]^2。 \tag{5.7}$$

当线性概率模型成立时,$X^2(L)$渐近服从自由度为 df $= I - 2$ 的卡方分布。它可以用来检验模型的拟合情况。统计量 z^2 的自由度为 df = 1,检验在式 5.6 模型中样本比例的线性趋势 $H_0:\beta = 0$。利用该统计量进行的独立性检验被称为 *Cochran-Armitage 趋势检验*(*Cochran-Armitage trend test*)。

以上分析似乎与线性 Logit 模型没有关系。不过,Cochran-Armitage 统计量与线性 Logit 模型中检验 $H_0:\beta = 0$ 的计分统计量是等价的。另外,Cochran-Armitage 统计量还与用于检验 $I \times J$ 表格中的线性趋势的统计量 M^2(式 3.15)有关;事实上,当 $J = 2$ 时,它等于 M^2,只不过用 n 代替了$(n-1)$。当 $I = 2$ 时,$X^2(L) = 0$ 且 $z^2 = X^2(I)$。

对于表 5.3 中有关饮酒与畸形的例子,$X^2(I) = 12.1$。沿用线性 Logit 模型中的赋值,Cochran-Armitage 趋势检验的结果为 $z^2 = 6.6$(P 值 = 0.010)。该检验给出了存在正向斜率的强有力的证据。另外,

$$X^2(I) = 12.1 = 6.6 + 5.5,$$

其中 $X^2(L) = 5.5$(df = 3)表明,样本比例对线性趋势的偏离很微弱。另外,趋势检验的结

果与基于原始数据($n=32\ 573$)的样本相关系数 $r=0.014$ 的 M^2(第3.4.5节)是一致的。就所选取的赋值而言,似乎只存在弱相关。然而,对于高度离散并且极不均衡的表格来说,r 作为一个描述关联强度的指标价值有限。

Cochran-Armitage 趋势检验(即计分检验)的结果通常与线性 Logit 模型中关于 $H_0:\beta=0$ 的沃尔德或似然比检验相似。当 $\{n_i\}$ 相等且 $\{x_i\}$ 采用等距赋值时,即使 n 很小,渐近结果也仍然成立。在表5.3中,沃尔德统计量等于 $(\hat{\beta}/\mathrm{SE})^2=(0.317/0.125)^2=6.4$ ($P=0.012$),似然比统计量等于4.25($P=0.039$)。由于表格中的计数极不均衡,这就意味着基于似然函数的似然比方法最为安全。这一点也同样适用于模型估计。表5.4给出的关于 β 的95%的剖面似然置信区间为(0.02,0.52),它要优于沃尔德区间 $0.317\pm1.96(0.125)=(0.07,0.56)$。即便这里的 n 非常大,第6.7.4节所介绍的小样本精确区间在此也是有价值的。

5.4 多元 Logistic 回归

与普通回归一样,logistic 回归也可以扩展到包括多个解释变量的情况。例如,对于 $\pi(\mathbf{x})=P(Y=1)$,具有 p 个预测变量 $\mathbf{x}=(x_1,\cdots,x_p)$ 的模型为

$$\mathrm{Logit}[\pi(\mathbf{x})]=\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p。\tag{5.8}$$

相应地,以 $\pi(\mathbf{x})$ 直接表示的公式为

$$\pi(\mathbf{x})=\frac{\exp(\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}{1+\exp(\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}。\tag{5.9}$$

其中,参数 β_i 可解释为在控制了其他的 x_j 后,x_i 对 $Y=1$ 的发生比的效应。相应地,$\exp(\beta_i)$ 表示在给定其他 x_j 的取值后,x_i 每增加1单位对发生比的可积效应。模型中的解释变量也可以是定性变量,通过虚拟变量来表示其类别。

5.4.1 多维列联表的 Logit 模型

当所有变量都是分类变量时,相应数据可以用多维列联表来表示。我们以二分预测变量 X 和 Z 来对此进行详细解释。我们将在 X 和 Z 的每个取值组合 (i,k) 处的样本规模视为给定的,且将每个取值组合上 Y 的两个计数视为相互独立的二项分布变量,用(0,1)表示每个变量的两个类别,并设 X 和 Z 的虚拟变量为 $x_1=z_1=1$ 以及 $x_2=z_2=0$。模型

$$\mathrm{Logit}[P(Y=1)]=\alpha+\beta_1x_i+\beta_2z_k\tag{5.10}$$

包括 X 和 Z 的主效应,但假定二者之间不存在交互效应,即一个因子的效应与另一个因子的取值无关。

在给定 Z 的取值 z_k 的情况下,改变 X 的类别对 Logit 的效应为

$$[\alpha+\beta_1(1)+\beta_2z_k]-[\alpha+\beta_1(0)+\beta_2z_k]=\beta_1。\tag{5.11a}$$

这个 Logit 的差等于对数发生比之差,也即在给定 Z 后 X 与 Y 之间的对数发生比之比。因而,$\exp(\beta_1)$ 是 X 和 Y 的条件发生比之比。在控制了 Z 后,当 $X=1$ 时成功的发生比是当 $X=0$ 时的 $\exp(\beta_1)$ 倍。这个条件发生比之比不会随着 Z 的变化而改变,也即,这里存*在同质的XY关联*(*homogeneous XY association*)(第2.3.5节)。在式5.10模型中不包括交互项意味着,所有分表都具有共同的发生比之比。当 $\beta_1=0$ 时,这个共同的发生比之比等于1。这时,在每个分表中 X 和 Y 相互独立,或者说*在给定Z的情况下二者条件独立*(*conditionally independent, given Z*)(第2.3.4节)。

Logit 尺度上的可加性，是分类变量之间不存在交互效应的一般定义。当然，也可以将其定义为其他尺度上的可加性，如 probit 或恒等连结。当在某一种尺度上存在显著的交互效应时，有可能在其他尺度上并不存在。在部分应用中，也许可以自然地选择某一特定的定义。例如，理论可能认为存在一个潜在的正态分布，并强调 probit 是各预测变量效应的可加函数。

如我们在第 5. 3. 2 节所示，包括 I 个类别的因子需要 $I-1$ 个虚拟变量。另一种对这类因子的表述与 ANOVA 模型中的表述相似。模型公式为

$$\mathrm{Logit}[P(Y=1)] = \alpha + \beta_i^X + \beta_k^Z, \tag{5.11b}$$

将 X 和 Z 的效应分别记作参数 $\{\beta_i^X\}$ 和 $\{\beta_k^Z\}$（这里 X 和 Z 的上标只是一种标识，并不代表乘方）。式 5. 11b 模型的形式适用于具有任意类别的 X 和 Z。参数 β_i^X 表示 X 的第 i 个类别对 Logit 的效应。在给定 Z 的情况下，X 和 Y 条件独立对应于 $\beta_1^X=\beta_2^X=\cdots=\beta_I^X$，此时，$P(Y=1)$ 不随 i 的变化而变化。

对于每个因子，式 5. 11b 模型中有一个参数是冗余的。将其中一个设定为 0，如 $\beta_I^X=\beta_K^Z=0$，表示这个类别本身不具有虚拟变量。当 X 和 Z 都只具有两个类别时，式 5. 11 模型的参数结构与式 5. 10 模型相对应，其中 $\beta_1^X=\beta_1$ 和 $\beta_2^X=0$，以及 $\beta_1^Z=\beta_2$ 和 $\beta_2^Z=0$。

5.4.2 例子：艾滋病与 AZT

表 5. 5 取自一项关于使用 AZT 对减缓艾滋病症状的效果的研究。在该研究中，338 名感染艾滋病病毒后免疫系统开始出现问题的退伍军人被随机分为两组，一组立即使用 AZT，另一组等到患者的 T 细胞显示出严重的免疫力衰退时才使用。表 5. 5 将这些研究对象按照种族、是否立即使用 AZT，以及在三年的跟踪期内是否出现了艾滋病症状进行了交叉分组。

在式 5. 10 模型中，我们用 X 表示 AZT 治疗（$x_1=1$ 表示立即使用 AZT，否则 $x_2=0$），Z 表示种族（$z_1=1$ 表示白人，$z_2=0$ 表示黑人），以此来预测出现艾滋病症状的概率。这样，α 就是种族为黑人的研究对象在没有立即使用 AZT 的情况下出现艾滋病症状的对数发生比，β_1 是那些立即使用 AZT 的对象的对数发生比的增量，以及 β_2 是研究对象为白人的对数发生比的增量。表 5. 6 给出了模型输出结果。关于立即使用 AZT 与出现艾滋病症状的对数发生比之比的估计值等于 $\exp(-0.7195)=0.487$。对于每个种族，立即使用 AZT 的对象出现症状的估计发生比约相当于那些没有立即使用的对象的一半。该效应的沃尔德置信区间为 $\exp[-0.720\pm1.96(0.279)]=(0.28,0.84)$。利用似然方法求得的区间与此相似。

表 5. 5　按照是否使用 AZT 和种族划分的出现艾滋病症状的情况

种族	使用 AZT	症状	
		是	否
白人	是	14	93
	否	32	81
黑人	是	11	52
	否	12	31

来源：《纽约时报》，1991 年 2 月 15 日。

在控制种族后,关于 AZT 治疗和出现艾滋病症状的条件独立性假设在式 5. 10 模型中可表示为 $H_0:\beta_1=0$。将式 5. 10 模型与具有 $\beta_1=0$ 的较简单模型进行比较的似然比统计量等于 6. 9(df = 1),这表明二者之间存在关联($P=0.01$)。相应的沃尔德统计量 $(\hat{\beta}_1/SE)^2=(-0.720/0.279)^2=6.65$ 也给出了相似的结果。

表 5. 6　对艾滋病症状数据拟合 Logit 模型的输出结果

Goodness-of-Fit Statistics

Criterion	DF	Value	Pr > ChiSq
Deviance	1	1.383 5	0.239 5
Pearson	1	1.391 0	0.238 2

Parameter	Estimate	Std Error	Wald Chi-Square	Pr > ChiSq
Intercept	-1.073 6	0.262 9	16.670 5	<.000 1
azt	-0.719 5	0.279 0	6.650 7	0.009 9
race	0.055 5	0.288 6	0.037 0	0.847 6

Odds Ratio Estimates

Effect	Estimate	95% Wald	Confidence Limits
azt	0.487	0.282	0.841
race	1.057	0.600	1.861

Profile Likelihood Confidence Interval for Odds Ratios

Effect	Estimate	95% Confidence	Limits
azt	0.487	0.279	0.835
race	1.057	0.605	1.884

Obs	race	azt	y	n	Pi_hat	lower	upper
1	1	1	14	107	0.149 62	0.098 97	0.219 87
2	1	0	32	113	0.265 40	0.196 68	0.347 74
3	0	1	11	63	0.142 70	0.087 04	0.225 19
4	0	0	12	55	0.254 72	0.169 53	0.363 96

表5. 7 给出了对式5. 11 模型的三种不同参数定义方式所对应的估计结果:①将最后一个参数设定为 0;②将第一个参数设定为 0;③将参数相加之和设定为零。无论使用哪一种方式,在给定使用 AZT 和种族的取值时,出现艾滋病症状的估计概率都相同。例如,

表 5. 7　对表 5. 5 数据拟合 Logit 模型的参数估计

参数	参数的定义		
	最后一个等于 0	第一个等于 0	相加等于 0
截距	-1.074	-1.738	-1.406
AZT 是	-0.720	0.000	-0.360
否	0.000	0.720	0.360
种族白人	0.055	0.000	0.028
黑人	0.000	-0.055	-0.028

在任一情况下,估计的截距,加上立即使用 AZT 的参数,再加上种族为白人的参数都是

-1.738,所以估计的白人退役军人在立即使用 AZT 的情况下出现艾滋病症状的概率等于 $\exp(-1.738)/[1+\exp(-1.738)]=0.15$。表 5.6 的底部给出了这些概率的点估计和区间估计。图 5.4 显示了相应的样本比例(图中的四个圆点)、点估计,以及 95% 的置信区间。

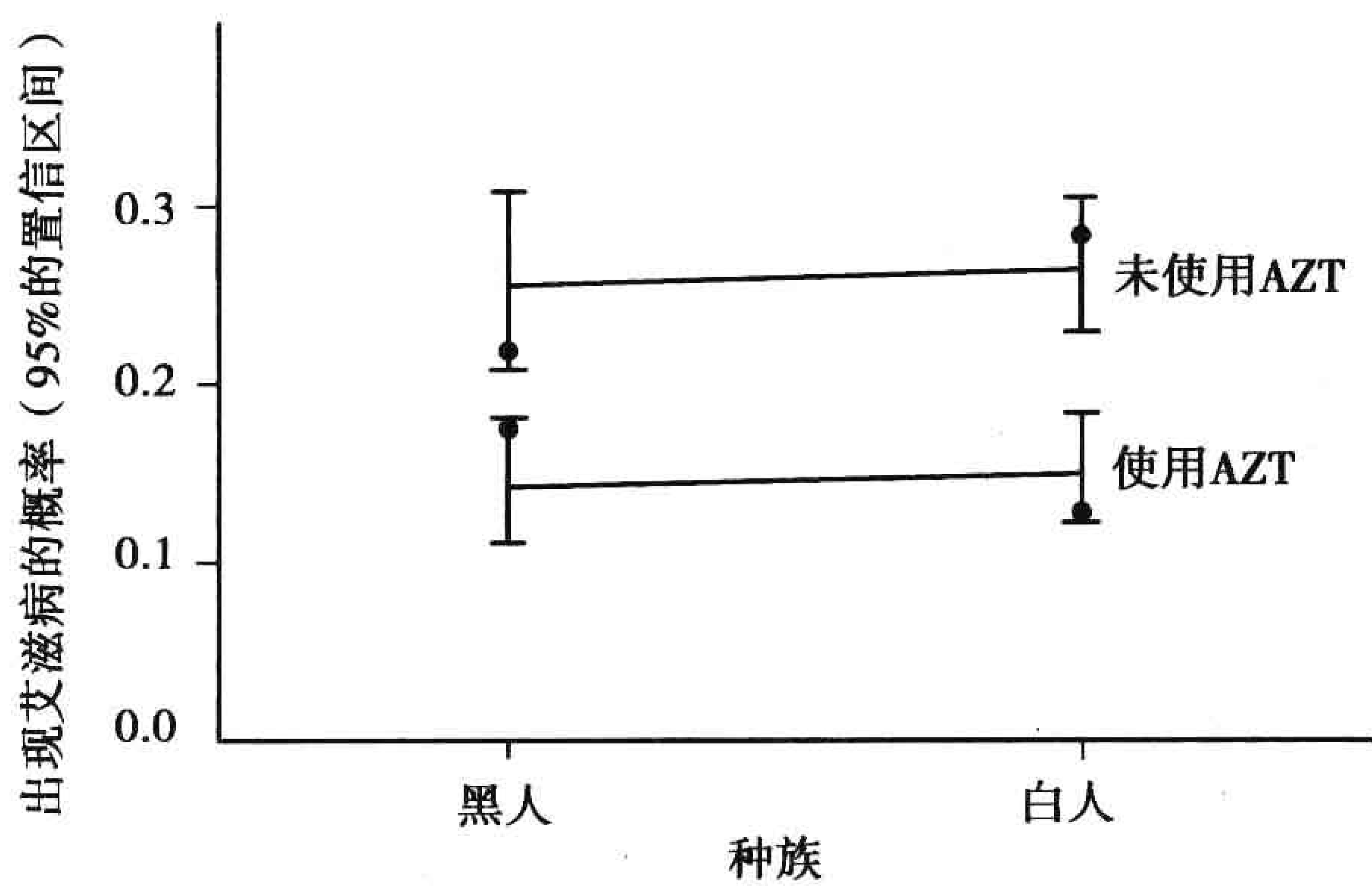


图 5.4　种族和使用 AZT 对出现艾滋病症状的概率估计值的影响(圆点表示样本比例)

类似地,对于每一种参数定义方法,$\beta_1^X-\beta_2^X$ 完全相等,它代表了在给定 Z 后 X 和结果变量的条件对数发生比之比。在这里,$\exp(\hat{\beta}_1^X-\hat{\beta}_2^X)=\exp(-0.720)=0.49$ 所估计的是不同种族在立即使用 AZT 与出现艾滋病症状之间共同的发生比之比。

5.4.3　作为似然比检验的拟合优度

通过比较拟合模型 M_1 对应的对数似然函数 L_1 与较简单模型 M_0 对应的对数似然函数 L_0,似然比统计量 $-2(L_0-L_1)$ 检验是否 M_1 中的某些参数等于零。记 $G^2(M_0 \mid M_1)$ 为在模型 M_1 成立的情况下关于模型 M_0 的检验统计量。拟合优度统计量 $G^2(M)$ 是当 $M_0=M$,M_1 为饱和模型时的特例。在检验 M 的拟合情况时,我们是在检验是否*所有*包括在饱和模型而未包括在模型 M 中的参数都等于零。检验的渐近自由度等于两个模型的参数个数之差,也即饱和模型中二项分布的数量减去模型 M 的参数个数。

我们以前面提到的艾滋病数据为例,对式 5.10 模型进行拟合优度检验。模型的拟合结果表明,白人退伍军人立即使用了 AZT 后在研究期间出现艾滋病症状的估计概率为 0.150。由于 107 名白人退伍军人使用了 AZT,出现症状的拟合计数为 $107(0.150)=16.0$,未出现症状的拟合值为 $107(0.850)=91.0$。类似地,读者可以求得表 5.5 中每个单元格的拟合值。将这些拟合值与单元格计数进行比较,则有模型的拟合优度统计量是 $G^2=1.38$ 和 $X^2=1.39$。模型包括四个二项分布,是否使用 AZT 和种族的每一取值组合各对应一个。由于该模型包括三个参数,残差自由度为 $\mathrm{df}=4-3=1$。G^2 和 X^2 的值很小,表明模型拟合得不错($P>0.2$)。

对于式 5.10 模型,X 和 Y 之间的发生比之比在 Z 的每一个取值上都是相同的。拟合优度检验对此结构进行了检查,也即,该检验也是对发生比之比是否具有同质性的检验。就表 5.5 来说,同质性是可能的。由于残差自由度等于 1,增加一个交互项而允许两个发生比之比不同的更复杂模型就是饱和模型。

令 L_S 表示饱和模型的最大对数似然值。如同第 4.5.4 节讨论的那样,比较模型 M_1 和 M_0 的似然比统计量是

$$G^2(M_0 \mid M_1)=-2(L_0-L_1)$$

$$= -2(L_0 - L_S) - [-2L_1 - L_S)]$$
$$= G^2(M_0) - G^2(M_1)。$$

比较两个模型的检验统计量与两个模型的拟合优度统计量（偏离度）G^2 之差完全相同。举例来说，考虑艾滋病数据中关于种族效应的假设 $H_0:\beta_2=0$。似然比统计量等于0.04，表明较简单的模型是充分的；同时，它等于 $G^2(M_0)-G^2(M_1)=1.42-1.38$，其中 M_0 是当 $\beta_2=0$ 时的较简单模型。

即使单个的 $G^2(M_i)$ 不服从卡方分布，用于模型比较的统计量通常近似服从卡方零分布。例如，当某一预测变量是连续变量或者列联表中存在非常小的拟合值时，$G^2(M_i)$ 的样本分布可能与卡方分布相去甚远。不过，除非用于模型比较的统计量具有很大的自由度（即所比较的两个模型相差很多个参数），$G^2(M_0|M_1)$ 的零分布近似于卡方分布。

5.4.4 例子：再论马蹄蟹数据

与普通回归一样，logistic 回归既可以包括定量的，也可以包括定性的预测变量。我们以马蹄蟹数据（第5.1.3节）为例，利用雌蟹的壳宽和颜色作为预测变量。颜色包括五个类别：浅色、浅褐色、褐色、深褐色、黑色，它反映的是蟹的年龄，年龄较大的蟹颜色一般会更暗一些。样本中没有浅色的蟹，因此，我们的模型只包括后四种类别的颜色。

首先，将颜色视为定性变量。四个类别可以用3个虚拟变量来表示。模型可记为

$$\text{Logit}(\pi) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x, \tag{5.12}$$

其中 $\pi=P(Y=1)$，$x=$ 以厘米为单位的壳宽，并且

对于浅褐色的蟹，$c_1 = 1$，否则为0，

对于褐色的蟹，$c_2 = 1$，否则为0，

对于深褐色的蟹，$c_3 = 1$，否则为0。

当 $c_1=c_2=c_3=0$ 时，蟹的颜色为黑色（第4类）。表5.8给出了参数的最大似然估计结果。例如，对于黑色的蟹，$\text{Logit}(\hat{\pi}) = -12.715+0.468x$；与之相比，对于浅褐色的蟹，有 $c_1=1$，则有 $\text{Logit}(\hat{\pi}) = (-12.715+1.330)+0.468x = -11.385+0.468x$。当壳宽等于其均值（26.3厘米）时，黑色蟹对应的 $\hat{\pi}=0.399$，而浅褐色蟹为0.715。

表5.8 以壳宽和颜色为预测变量的模型输出结果

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	168	187.457 0
Pearson Chi-Square	168	168.659 0
Log Likelihood		-93.728 5

Parameter	Estimate	Standard Error	Likelihood-Ratio 95% Confidence	Limits	Chi-Square	Pr > ChiSq
intercept	-12.715 1	2.761 8	-18.456 4	-7.578 8	21.20	<.000 1
c1	1.329 9	0.852 5	-0.273 8	3.135 4	2.43	0.118 8
c2	1.402 3	0.548 4	0.352 7	2.526 0	6.54	0.010 6
c3	1.106 1	0.592 1	-0.027 9	2.313 8	3.49	0.061 7
width	0.468 0	0.105 5	0.271 3	0.687 0	19.66	<.000 1

模型假定颜色和壳宽的影响不存在交互项，即对于所有的颜色而言，壳宽的系数是

相同的(0.468),反映壳宽和 π 的关系的曲线形状也完全一样。对于每一种颜色,壳宽每增加 1 厘米,都对 $Y=1$ 的发生比产生一个 exp(0.468) =1.60 的可积效应。图 5.5 展示了所拟合的模型。任意两条曲线的形状都相同,每一条曲线都可以通过对另一条曲线的左右平移来得到。

曲线在水平方向上相互平行意味着,任何两条曲线都不会相交。对于所有的壳宽取值,第四类颜色(黑色)拥有同伴的估计概率都比其他颜色低。在这里,壳宽存在着明显的正效应。

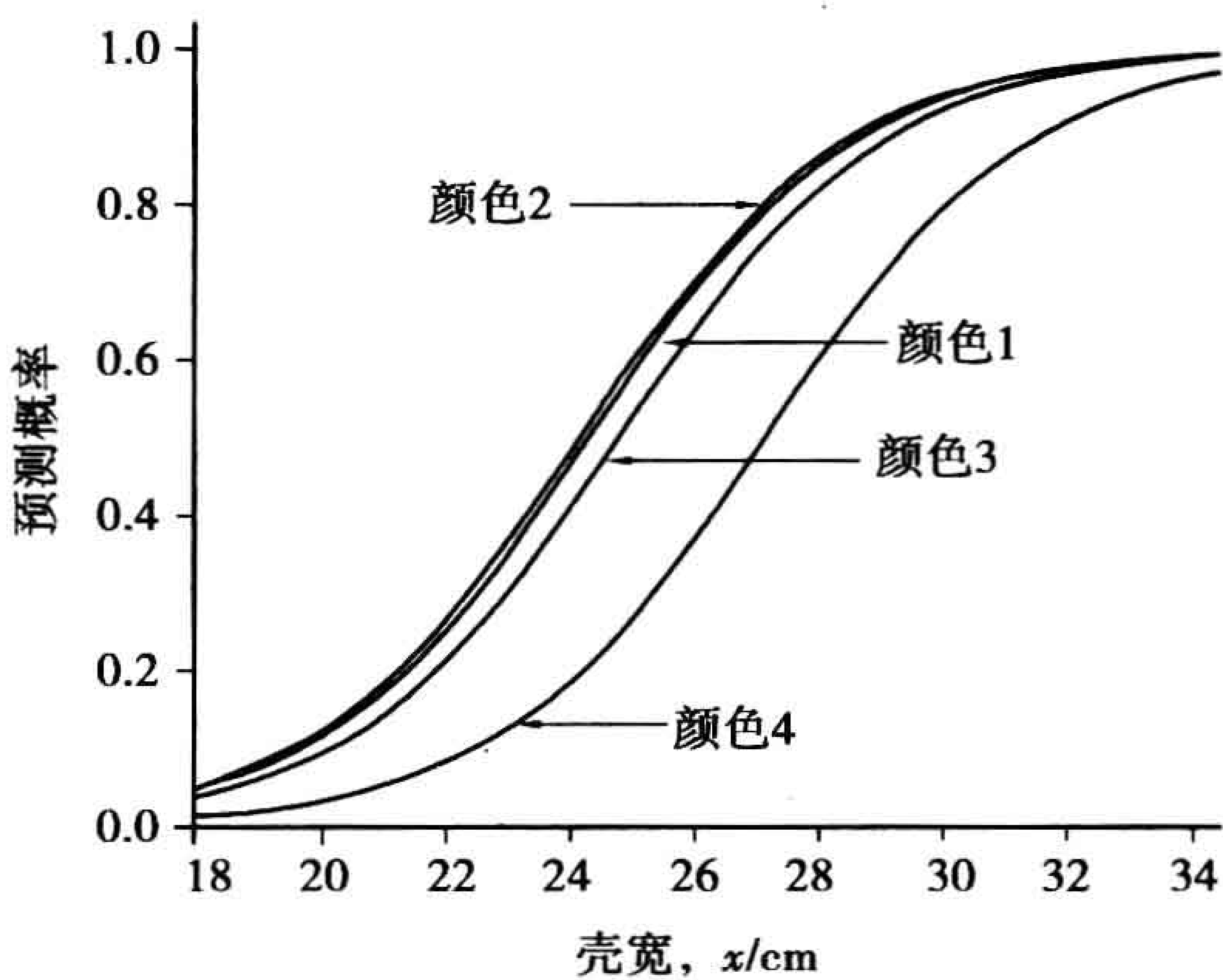


图 5.5 以壳宽和颜色为预测变量对马蹄蟹是否拥有同伴拟合的 Logistic 回归模型

对两种颜色的参数估计值之差求幂便可得这两种颜色的发生比之比。例如,对于浅褐色和黑色的蟹,其参数估计值之差等于 1.330。因而,在任意给定的壳宽水平上,一只浅褐色蟹拥有同伴的估计发生比是黑色蟹的 exp(1.330) =3.8 倍。当壳宽 $x=26.3$ 时,浅褐色蟹的发生比等于 0.715/0.285 =2.51,而黑色蟹的发生比仅为 0.399/0.601 = 0.66,二者的发生比之比等于 2.51/0.66 =3.8。

5.4.5 模型比较

为了检验颜色对式 5.12 模型的贡献是否显著,我们检验假设 $H_0: \beta_1=\beta_2=\beta_3=0$。这个假设意味着,在控制了壳宽之后,拥有同伴的概率独立于颜色。我们将式 5.12 完全模型的最大对数似然值 L_1 与较简单模型的最大对数似然值 L_0 进行对比。检验统计量 $-2(L_0-L_1)=7.0$,具有 df =3 个自由度,即两个模型的参数数量之差。卡方检验的 P 值为 0.07,即有一定证据表明颜色效应是存在的。

允许颜色和壳宽的交互效应的更复杂模型多出来三项,即壳宽与代表颜色的虚拟变量的交叉乘积项。这个模型等价于对每种颜色的蟹分别拟合以壳宽为预测变量的 logistic 回归模型。这时,每种颜色所对应的反映壳宽和 $P(Y=1)$ 的关系的曲线具有不同的形状,所以对于两种颜色的比较会随着壳宽的不同取值而变化。比较包括和不包括交互项的模型的似然比统计量等于 4.4,自由度为 df =3。因此,没有证据表明存在交互效应($P=0.22$)。

5.4.6 将定序预测变量视为定量变量

从最浅到最深,颜色的类别是可以排序的。一种简单的模型将这一变量视为定量变量。利用一组单调的赋值,颜色的效应可能是线性的。具体来说,利用赋值 $c=\{1,2,3,4\}$ 来代表颜色的类别,模型可表示为

$$\text{Logit}(\pi) = \alpha + \beta_1 c + \beta_2 x, \tag{5.13}$$

其拟合结果为$\hat{\beta}_1 = -0.509(\text{SE}=0.224)$，$\hat{\beta}_2 = 0.458(\text{SE}=0.104)$。这表明，每个变量都存在显著的效应。对于给定的壳宽，颜色深度每增加一个类别，拥有同伴的估计发生比将变化 $\exp(-0.509)=0.60$ 倍。

将该模型与每种颜色都具有单独参数的更复杂模型式 5.12 进行比较，似然比统计量等于 1.7(df=2)。这个检验考察的是，在式 5.12 模型成立的条件下，较简单的模型拟合是否充分，也即，当将上述颜色的赋值与式 5.12 模型中颜色的参数进行绘图时，这些参数是否具有线性趋势。检验结果表明，上述模型的简化是可以接受的($P=0.44$)。

在将颜色作为定性变量的式 5.12 模型中，颜色所对应的参数估计值为(1.33,1.40,1.11,0)，其中黑色类别所对应的 0 值反映了它为参照组。尽管这些值并没有明显地偏离线性趋势，但是与最后一个值相比，前三个值都很接近。因此，式 5.13 模型中对颜色的另一种可能赋值为(1,1,1,0)，也即，对于黑色的蟹，赋值为 0，其他赋值为 1。将使用这些二分赋值的式 5.13 与式 5.12 模型进行比较，似然比统计量等于 0.5(df=2)，表明这个较简单模型的拟合也是充分的。该模型的拟合结果为

$$\text{Logit}(\hat{\pi}) = -12.980 + 1.300c + 0.478x, \tag{5.14}$$

模型参数对应的标准误分别为 0.526 和 0.104。在给定壳宽后，颜色较浅的蟹拥有同伴的估计发生比是黑色蟹的 $\exp(1.300)=3.7$ 倍。

由以上分析可见，将颜色作为定性变量、定量变量并使用赋值(1,2,3,4)、二分变量并使用赋值(1,1,1,0)的模型均表明，黑色蟹拥有同伴的可能性最小。要想确定哪一种颜色赋值最恰当，则需要一个更大规模的样本。当将定序的预测变量作为定量变量使用且模型拟合很好时，这样做是有价值的。这时，模型更加简单，解释起来也更容易；并且当预测变量只对应一个参数而不是几个参数时，对其效应的检验也更具统计效能。关于这一问题，我们将在第 6.4 节进一步讨论。

5.4.7 标准化系数与基于概率的解释

为了比较具有不同度量单位的定量预测变量的效应大小，报告标准化的系数很有意义。一种方法是使用标准化的预测变量来拟合模型，即将 x_j 替代为$(x_j - \bar{x}_j)/s_{x_j}$。这时，每个回归系数表示的是在控制其他变量后，预测变量每变化一个标准差所导致的效应。等价地，对于每个 j，大家也可以将非标准化的估计值 $\hat{\beta}_j$ 乘上 s_{x_j}(另见注解 5.9)。

不论度量单位是什么，许多人不能理解发生比或发生比之比的效应。对模型进行线性化处理(第 5.1.1 节)，将模型解释为概率的相应变动的方法较为简单，同时也适用于多个预测变量的情形。考虑满足 $\hat{P}(Y=1)=\hat{\pi}$ 的一组预测变量的取值。这时，在控制其他预测变量的情况下，x_j 每增加 1 单位所对应的 $\hat{\pi}$ 的变化大约为 $\hat{\beta}_j\hat{\pi}(1-\hat{\pi})$。例如，在式 5.14 模型的拟合结果中，当 $\hat{\pi}=0.5$ 时，壳宽每增加 1 厘米的效应大约为$(0.478)(0.5)(0.5)=0.12$。这个效应很大，因为 1 厘米的变动还不到壳宽的半个标准差。

随着预测变量取值变动的增加，这种线性近似变得不准确。更精确的解释是直接使用概率公式。为了描述 x_j 的效应，可以将其他预测变量的取值设定为样本均值，然后计算在 x_j 取最小值和最大值时分别对应的估计概率。不过，这种方法受奇异值影响很大。通常来说，使用四分位点更为合理一些。

对于式 5.14 模型的拟合结果，x 和 c 的样本均值分别为 26.3 和 0.873。x 的第一个和第三个四分位点分别是 24.9 和 27.7。当 $x=24.9$ 且 $c=\bar{c}$ 时，$\hat{\pi}=0.51$；当 $x=27.7$ 且

$c=\bar{c}$时，$\hat{\pi}=0.80$。在壳宽中间50%的取值范围内，$\hat{\pi}$ 从0.51 增加到0.80，这表明壳宽存在很强的效应。由于c的取值仅为0 和1，因而，可以分别报告在c的每个取值下的效应。另外，当某一解释变量为虚拟变量时，报告在它的两个取值处的估计概率比四分位点更有意义，而且四分位点可能会取相同的值。在 $\bar{x}=26.3$ 的情况下，当 $c=0$ 时 $\hat{\pi}=0.40$，当 $c=1$ 时 $\hat{\pi}=0.71$。区分黑色蟹和其他蟹的颜色效应同样也很大。

表5.9 给出了一种报告效应的方法，它可以让那些不熟悉发生比之比的人也能够理解。该表还给出了对式5.14 模型的扩展——允许存在交互项的结果。这时，估计的壳宽效应对于较浅色的蟹更强。但是，交互项本身并不显著。

表5.9　用壳宽和颜色来预测是否存在同伴的式5.14 模型的效应

变　量	估计值	标准误	比　较	概率的变动
不包括交互项的模型				
截距	-12.980	2.727		
颜色(0=黑色，1=其他)	1.300	0.526	(1,0) at $\bar{x}$	0.31=0.71-0.40
壳宽，x/cm	0.478	0.104	(UQ,LQ) at $\bar{c}$	0.29=0.80-0.51
包括交互项的模型				
截距	-5.854	6.694		
颜色(0=黑色，1=其他)	-6.958	7.318		
壳宽，x/cm	0.200	0.262	(UQ,LQ) at $c=0$	0.13=0.43-0.30
壳宽×颜色	0.322	0.286	(UQ,LQ) at $c=1$	0.29=0.84-0.55

5.5　Logistic 回归模型的拟合

Logistic 回归的最大似然估计与模型拟合是第4.6 节所介绍的广义线性模型拟合过程的一个特例。给定 n 个对象，我们将其视为 n 个独立的二分结果变量。令 $\mathbf{x}_i=(x_{i1},\cdots,x_{ip})$表示 p 个解释变量的第 i 种取值组合，其中 $i=1,\cdots,N$。当解释变量是连续变量时，每个对象都可能对应着一种不同的取值，这时 $N=n$。将 α 视作具有单位系数的回归参数，logistic 回归模型(式5.8)可表述为：

$$\pi(x_i)=\frac{\exp(\sum_{j=1}^{p}\beta_j x_{ij})}{1+\exp(\sum_{j=1}^{p}\beta_j x_{ij})}。\tag{5.15}$$

5.5.1　似然方程

当某一给定的 x_i 值处存在多个观测值时，记录相应的观测值数量 n_i 以及成功次数就足够了。令 y_i 表示成功的计数而不是单个的二分结果变量，这时，$\{Y_1,\cdots,Y_N\}$ 服从独立的二项分布，并且 $E(Y_i)=n_i\pi(x_i)$，其中 $n_1+\cdots+n_N=n$。它们的联合概率密度函数与 N 个二项分布函数的乘积成比例，

$$\prod_{i=1}^{N}\pi(\mathbf{x}_i)^{y_i}[1-\pi(\mathbf{x}_i)]^{n_i-y_i}$$

$$= \left\{ \prod_{i=1}^{N} \exp\left[\log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right)^{y_i} \right] \right\} \left\{ \prod_{i=1}^{N} [1 - \pi(\mathbf{x}_i)]^{n_i} \right\}$$

$$= \left\{ \exp\left[\sum_i y_i \log \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right] \right\} \left\{ \prod_{i=1}^{N} [1 - \pi(\mathbf{x}_i)]^{n_i} \right\}。$$

对于式 5.15 模型，第 i 个 Logit 等于 $\sum_j \beta_j x_{ij}$，所以上式中最后一个表达式的指数项等于 $\exp\left[\sum_i y_i (\sum_j \beta_j x_{ij}) \right] = \exp\left[\sum_j (\sum_i y_i x_{ij}) \beta_j \right]$。同时，由于 $[1 - \pi(x_i)] = \left[1 + \exp(\sum_j \beta_j x_{ij})\right]^{-1}$，对数似然函数等于

$$L(\boldsymbol{\beta}) = \sum_j (\sum_i y_i x_{ij}) \beta_j - \sum_i n_i \log\left[1 + \exp(\sum_j \beta_j x_{ij})\right]。 \quad (5.16)$$

它仅通过充分统计量 $\left\{ \sum_i y_i x_{ij}, \quad j = 1, \cdots, p \right\}$ 取决于二项分布计数。

设 $\partial L(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = 0$ 便得到似然方程。由于

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i y_i x_{ij} - \sum_i n_i x_{ij} \frac{\exp(\sum_k \beta_k x_{ik})}{1 + \exp(\sum_k \beta_k x_{ik})},$$

似然方程为

$$\sum_i y_i x_{ij} - \sum_i n_i \hat{\pi}_i x_{ij} = 0, \quad j = 1, \cdots, p, \quad (5.17)$$

其中 $\hat{\pi}_i = \exp(\sum_k \hat{\beta}_k x_{ik}) / \left[1 + \exp(\sum_k \hat{\beta}_k x_{ik})\right]$ 是对 $\pi(\mathbf{x}_i)$ 的最大似然估计。不难看出，这些方程式是式 4.25 中二项分布广义线性模型的方程式的一个特例（在那里，y_i 是指成功的比例）。这些方程是非线性的，因而需要通过迭代法求解。

令 $\mathbf{X}$ 表示 $\{x_{ij}\}$ 的 $N \times p$ 矩阵，式 5.17 似然方程可以表示为

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\boldsymbol{\mu}}, \quad (5.18)$$

其中 $\hat{\mu}_i = n_i \hat{\pi}_i$。这个方程展示了一个基本结论：对于具有典型连结的广义线性模型，其似然方程就是使所估计的期望值等于其充分统计量。方程式 4.44 在广义线性模型的框架下显示了这一结果，而方程式 5.18 则是普通回归中的正规方程。

5.5.2 参数估计的渐近协方差矩阵

最大似然估计 $\hat{\boldsymbol{\beta}}$ 服从大样本正态分布，并且它的协方差矩阵等于信息矩阵的逆矩阵。观察到的信息矩阵的元素为

$$-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_a \partial \beta_b} = \sum_i \frac{x_{ia} x_{ib} n_i \exp(\sum_j \beta_j x_{ij})}{[1 + \exp(\sum_j \beta_j x_{ij})]^2} = \sum_i x_{ia} x_{ib} n_i \pi_i (1 - \pi_i)。 \quad (5.19)$$

它不是 $\{y_i\}$ 的函数，所以观察信息矩阵和期望的信息矩阵完全一样。这种情况对于所有使用典型连结的广义线性模型都成立（第 4.6.4 节）。

估计的协方差矩阵是具有元素如式 5.19 的矩阵的逆矩阵，并代入 $\hat{\boldsymbol{\beta}}$，其形式为

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \{\mathbf{X}' \mathbf{diag}[n_i \hat{\pi}_i (1 - \hat{\pi}_i)] \mathbf{X}\}^{-1} \quad (5.20)$$

其中 $\mathbf{diag}[n_i \hat{\pi}_i (1 - \hat{\pi}_i)]$ 表示主对角线元素为 $\{n_i \hat{\pi}_i (1 - \hat{\pi}_i)\}$ 的 $N \times N$ 的对角矩阵。这是广义线性模型的协方差矩阵（式 4.28）的一个特例，其中所估计的对角加权矩阵 $\hat{\mathbf{W}}$ 的元素为 $\hat{w}_i = n_i \hat{\pi}_i (1 - \hat{\pi}_i)$。式 5.20 矩阵中主对角线元素的平方根是对 $\hat{\boldsymbol{\beta}}$ 的标准误的估计。

5.5.3 概率估计值的分布

通过$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$,可以对 $\boldsymbol{\beta}$ 以及相应的效应(如发生比之比)进行统计推断,也可以构建在 $\mathbf{x}$ 的特定取值处结果变量的概率 $\pi(\mathbf{x})$ 的置信区间。

关于 $\text{Logit}[\hat{\pi}(\mathbf{x})] = \mathbf{x}\hat{\boldsymbol{\beta}}$ 的方差估计等于 $\mathbf{x}\,\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\,\mathbf{x}'$。在大样本的情况下,$\text{Logit}[\hat{\pi}(\mathbf{x})] \pm z_{\alpha/2}\sqrt{\mathbf{x}\,\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}'}$ 是真实 Logit 的置信区间。利用 $\pi = \exp(\text{Logit})/[1+\exp(\text{Logit})]$ 对区间端点进行转换,便可求得 $\pi(\mathbf{x})$ 的置信区间。

5.5.4 Newton-Raphson 法在 Logistic 回归中的应用

现在回到第 4.6.1 节关于 Newton-Raphson 迭代法的介绍。令

$$u_j^{(t)} = \left.\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}^{(t)}} = \sum_i (y_i - n_i\pi_i^{(t)})x_{ij},$$

$$h_{ab}^{(t)} = \left.\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_a \partial \beta_b}\right|_{\boldsymbol{\beta}^{(t)}} = -\sum_i x_{ia}x_{ib}n_i\pi_i^{(t)}(1-\pi_i^{(t)})。$$

这里,$\boldsymbol{\pi}^{(t)}$,即对 $\hat{\boldsymbol{\pi}}$ 的第 t 次近似,可以通过以下公式由 $\boldsymbol{\beta}^{(t)}$ 来求得:

$$\pi_i^{(t)} = \frac{\exp\left(\sum_{j=1}^{p}\beta_j^{(t)}x_{ij}\right)}{1+\exp\left(\sum_{j=1}^{p}\beta_j^{(t)}x_{ij}\right)}。\tag{5.21}$$

利用公式 4.39 中的 $\mathbf{u}^{(t)}$ 和 $\mathbf{H}^{(t)}$ 来求下一个 $\boldsymbol{\beta}^{(t+1)}$,

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \{\mathbf{X}'\mathbf{diag}[n_i\pi_i^{(t)}(1-\pi_i^{(t)})]\mathbf{X}\}^{-1}\mathbf{X}'(\mathbf{y}-\boldsymbol{\mu}^{(t)}),\tag{5.22}$$

其中 $\mu_i^{(t)} = n_i\pi_i^{(t)}$。再用 $\boldsymbol{\beta}^{(t+1)}$ 求 $\boldsymbol{\pi}^{(t+1)}$,依此类推。

对于初始猜测 $\boldsymbol{\beta}^{(0)}$,公式 5.21 给出了 $\boldsymbol{\pi}^{(0)}$,对于 $t>0$,利用式 5.22 和式 5.21 进行如上式所示的迭代求解。$\boldsymbol{\pi}^{(t)}$ 和 $\boldsymbol{\beta}^{(t)}$ 的极限值收敛于最大似然估计值 $\hat{\boldsymbol{\pi}}$ 和 $\hat{\boldsymbol{\beta}}$(Walker and Duncan ,1967)。矩阵 $\mathbf{H}^{(t)}$ 收敛于 $\hat{\mathbf{H}} = -\mathbf{X}'\mathbf{diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)]\mathbf{X}$。根据式 5.20,Newton-Raphson 法也同时给出对 $\hat{\boldsymbol{\beta}}$ 的渐近协方差矩阵的估计,即 $-\hat{\mathbf{H}}^{-1}$。

由第 4.6.3 节的讨论可知,$\boldsymbol{\beta}^{(t+1)}$ 具有迭代再加权最小二乘法的形式 $(\mathbf{X}'\mathbf{V}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_t^{-1}\mathbf{z}^{(t)}$,其中 $\mathbf{z}^{(t)}$ 的元素为

$$z_i^{(t)} = \log\frac{\pi_i^{(t)}}{1-\pi_i^{(t)}} + \frac{y_i - n_i\pi_i^{(t)}}{n_i\pi_i^{(t)}(1-\pi_i^{(t)})},\tag{5.23}$$

且 $\mathbf{V}_t$ 是元素为 $\{1/n_i\pi_i^{(t)}(1-\pi_i^{(t)})\}$ 的对角矩阵。在这个表达式中,$\mathbf{z}^{(t)}$ 是样本数据中线性形式的 Logit 连结函数在 $\boldsymbol{\pi}^{(t)}$ 处的取值(参见式 4.42)。由第 3.1.6 节可知,$\mathbf{V}_t$ 的元素是对样本 Logit 的渐近方差的估计。最大似然估计等于一系列加权最小二乘法估计的极限,其中每一步所使用的权数矩阵都各不相同。

5.5.5 有限估计值的收敛性和存在性

Logistic 回归模型的对数似然函数是严格的凹函数。除某些边界上的情况以外,最大似然估计存在并且唯一(Haberman, 1974a; Wedderburn, 1976; Albert and Anderson, 1984)。当 $y=0$ 所对应的解释变量的取值与 $y=1$ 所对应的取值没有交叉时,最大似然估计不存在或者等于无穷大,也即,当一个超平面穿过由预测变量的值所形成的空间时,超平面的一边是所有 $y=0$ 的观测值,而另一边总是 $y=1$ 的观测值。这时存在着完全判别

(perfect discrimination),即大家根据预测变量的取值(除了可能落在边界上的点)可以完全地预测样本的结果。当二者存在重合时,最大似然估计存在并且唯一。Probit 和其他一些连结函数也具有类似的特征(Silvapulle,1981)。

图 5.6 展示了只有一个解释变量的情况。这里,在 $x=10,20,30,40$ 时,$y=0$,而在 $x=60,70,80,90$ 时,$y=1$。对此,理想的拟合结果是对于 $x\leqslant40$ 有 $\hat{\pi}=0$,以及对于 $x\geqslant60$ 有 $\hat{\pi}=1$。当 $\hat{\beta}\to\infty$ 时,对于给定的 $\hat{\beta}$ 使 $\hat{\alpha}=-\hat{\beta}(50)$,这样,在 $x=50$ 时 $\hat{\pi}=0.5$,由此便生成了一个逐渐接近于完全拟合的持续上升的似然值序列。

在实际应用中,许多软件无法识别 $\hat{\beta}=\infty$。在几步迭代拟合之后,对数似然函数在当前的估计值上呈扁平状,收敛条件得以满足。因为对数似然函数如此扁平,且方差来自于负二阶导数矩阵的逆矩阵,统计软件往往报告极大的标准误。例如,对于上述数据,SAS 的 PROC GENMOD 给出的结果为 $\text{Logit}(\hat{\pi})=-192.2+3.8x$,标准误分别为 8.0×10^8 和 1.5×10^7。

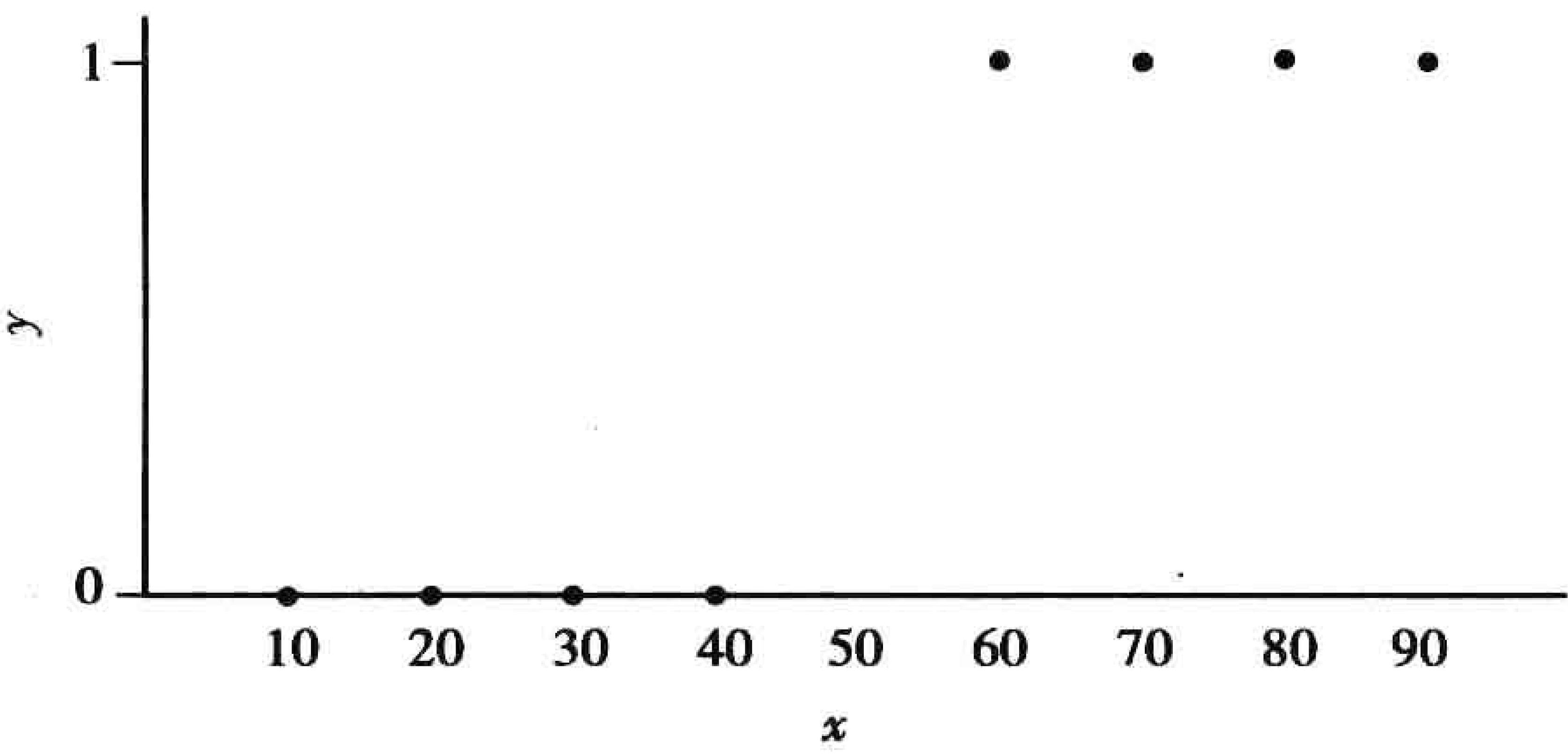


图 5.6 导致 logistic 回归参数估计值等于无穷大的完全判别

注 解

第 5.1 节:Logistic 回归参数的解释

5.1 专门介绍 logistic 回归应用的书包括:Collect(1991)、Hosmer and Lemeshow(2000)。书中包含相当的篇幅介绍 logistic 回归的有:Christensen(1997)、Cox and Snell(1989)、Morgan(1992)。Prentice(1976b)和 Stukel(1988)通过引入形状参数(shape parameters)对 logistic 回归进行了扩展,以修正在极端概率区域曲线的特性并允许两个尾部可以不对称。

5.2 Haldane(1956)建议将样本 Logit 的分子和分母分别加上 1/2。对于 n_i 较大的情况,这种调整所产生的偏差仅约为 $1/n_i^2$ 阶(参见:Firth (1993a),习题 14.4)。

5.3 Cornfield(1962)指出$(X|Y=i)$服从正态分布意味着 $P(Y=1|x)$ 为 logistic 曲线。这表明,logistic 回归可以应用于判别(discrimination)和分类(classification)分析,即利用研究对象的 x 值来预测其属于两个总体中的哪一个。Anderson(1975)、Bull and Donner(1987)、Efron(1975),以及 Press and Wilson(1978)通过比较,认为 logistic 回归优于判别分析,后者假定解释变量在每一个 Y 的取值上都服从正态分布。

5.4 Rosenbaum and Rubin(1983)利用 logistic 回归来修正在观察研究中比较两个组别时的偏差。他们将倾向(propensity)定义为给定一组解释变量 x 的取值,研究对象属于某一类别的概率。利用 logistic 回归,他们估计倾向是如何取决于 $\mathbf{x}$ 的取值的。他们证明,在比较结果变量的不同类别时,可以通过调整所估计的倾向来控制在 $\mathbf{x}$ 处各个类别的分布差异。这可以通过利用倾向来匹配各类样本或者按照倾向计分(propensity scores)的区间将研究对象进一步分层,以及将倾向包括到模型中进行

直接调整来实现。具体的做法,参见:D'Agostino(1998)。

5.5 Adelbasit and Plackett (1983)、Chaloner and Larntz (1988)、Minkin (1987),以及 Wu (1985)探讨了关于二分结果变量的实验设计问题,如选取预测变量的值以最优化估计参数值的条件或者估计满足结果变量的概率等于某个给定值时的预测变量取值。由于方差的变动性,这种做法具有一定难度。

第 5.2 节:Logistic 回归的统计推断

5.6 讨论了关于 logistic 回归的最大似然估计有:Albert and Anderson(1984),Berkson(1951,1953,1955),Cox(1958a),Hodges(1958),Walker and Duncan(1967)。对于复杂抽样调查的调整,参见:Hosmer and Lemeshow(2000,Sec. 6.4)、La Vange et al. (2001)。Scott 和 Wild (2001)讨论了复杂抽样设计下对个案—控制研究的分析。

5.7 Tsiatis(1980)提出了另一种拟合优度检验,将解释变量的值分割为一组区域,并且将代表各个区域的虚拟变量加入模型。检验统计量通过比较这个模型和较简单模型的拟合结果,检验是否有必要包括额外的参数。这种通过对解释变量的值进行分组,进而对比观察计数与拟合计数来检查模型拟合的思路,可以扩展到任何广义线性模型(Pregibon,1982)。Hosmer 等(1997)对各种不同做法进行了比较。

第 5.3 节:包括分类预测变量的 Logit 模型

5.8 对于 $P(Y=1)$ 的线性和 logistic 模型,Cochran-Armitage 趋势检验都具有局部渐近有效性。它相对于线性趋势的有效性是由于样本比例的近似正态性,当 $\beta=0$ 时对应的伯努利方差恒定。对于式 5.5 线性 Logit 模型,它的有效性源于其与计分检验的等价性。有关讨论,参见习题 9.35 以及:Cox(1958a)。Tarone 和 Gart(1980)表明,二分的线性趋势模型的计分检验不取决于连结函数。Gross(1981)指出,对于线性 Logit 模型,利用一组不正确的赋值进行独立性检验的局部渐近相对有效性等于真实赋值和不正确赋值之间的皮尔逊相关系数的平方。Simon(1978)给出了相应的渐近结果。Corcoran 等(2001)、Mantel(1963),以及 Podgor 等(1996)对趋势检验进行了扩展。

第 5.4 节:多元 Logistic 回归

5.9 由于标准化的 logistic 累积分布函数的标准差等于 $\pi/\sqrt{3}$,一些软件(如 SAS 中的 PROC LOGISTIC)将标准化系数定义为非标准化的回归系数乘以 $s_{x_j}\sqrt{3}/\pi$。

习题

应用部分

5.1 某项研究通过 logistic 回归来确定癌症患者中与症状缓和有关的因素,表 5.10 给出了最重要的解释变量,即标识指数(labeling index,LI),该指数测量在病人注射氚化胸腺核(tritiated thymidine)后细胞的增殖能力,代表被"标识"细胞所占的百分比。结果变量 Y 度量病人的症状是否得到缓解(1 = 是)。表 5.11 给出了利用 LI 来预测缓解的概率的 logistic 回归模型的输出结果。

a. 指出当 LI = 8 时,$\hat{\pi}=0.068$ 是如何得出的。

b. 证明当 LI = 26.0 时,$\hat{\pi}=0.5$。

c. 证明当 LI = 8 时,$\hat{\pi}$ 的变动率为 0.009;当 LI = 26.0 时,$\hat{\pi}$ 的变动率为 0.036。

d. 关于 LI 的第一和第三个四分位点分别为 14 和 28。证明在这些值之间,$\hat{\pi}$ 上升了 0.42,即从 0.15 增加到 0.57。

e. 证明对于 LI 每变动一个单位,估计的出现缓解的发生比变动 1.16 倍。

f. 说明如何获得输出结果中所给出的关于发生比之比的置信区间,加以解释。

表 5.10　习题 5.1 的数据

标识指数	观测值数量	出现缓解的数量	标识指数	观测值数量	出现缓解的数量	标识指数	观测值数量	出现缓解的数量
8	2	0	18	1	1	28	1	1
10	2	0	20	3	2	32	1	0
12	3	0	22	2	1	34	1	1
14	3	0	24	1	0	38	3	2
16	3	0	26	1	1			

来源:经授权重印自:E. T. Lee, *Comput. Prog. Biomed.* 4:80-92(1974)。

g. 构建关于该效应的沃尔德检验,加以解释。

h. 构建关于该效应的似然比检验,给出如何使用输出结果中的 $-2\log L$ 来构建检验统计量。

i. 说明如何求得在 LI = 8 时关于 π 的置信区间(*提示*:利用输出结果中的协方差矩阵)。

表 5.11　习题 5.1 的电脑输出结果

Criterion	Intercept Only	Intercept and Covariates
-2 Log L	34.372	26.073

Testing Global Null Hypothesis: BETA = 0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.298 8	1	0.004 0
Score	7.931 1	1	0.004 9
Wald	5.959 4	1	0.014 6

Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	-3.777 1	1.378 6	7.506 4	0.006 1
li	0.144 9	0.059 3	5.959 4	0.014 6

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
li	1.156	1.029	1.298

Estimated Covariance Matrix

Variable	Intercept	li
Intercept	1.900 616	-0.076 53
li	-0.076 53	0.003 521

Obs	li	remiss	n	pi_hat	lower	upper
1	8	0	2	0.067 97	0.011 21	0.319 25
2	10	0	2	0.088 79	0.018 09	0.340 10

译者注——li:标识指数。

5.2　在 1986 年挑战号发生灾难之前航天飞船的 23 次飞行中,表 5.12 给出了飞行时点

的温度以及是否至少有一个主要的 O 型圈出现热损坏的数据。

a. 利用 logistic 回归,构建温度对热损坏出现概率的效应的模型。将拟合的模型进行绘图,并加以解释。

b. 估计在华氏 31 度时,即挑战者号飞行的地点和时间的温度,出现热损坏的概率。

c. 构建温度对热损坏的效应的置信区间,并检验该效应的显著性。

d. 通过将该模型与更复杂的模型进行比较,检查模型拟合的情况。

表 5.12 习题 5.2 的数据[a]

Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD
1	66	0	2	70	1	3	69	0	4	68	0	5	67	0
6	72	0	7	73	0	8	70	0	9	57	1	10	63	1
11	70	1	12	78	0	13	67	0	14	53	1	15	67	0
16	75	0	17	70	0	18	81	0	19	76	0	20	79	0
21	75	1	22	76	0	23	58	1						

a Ft,飞行数;Temp,温度(°F);TD,热损坏(1,是;0,否)。

来源:数据来自:Table 1 in *J. Amer. Statist. Assoc.* ,84:945-957,(1989),by S. R. Dalal, E. B. Fowlkes and B. Hoadley。授权重印自:the *Journal of the American Statistical Association*。

5.3 参见表 4.2。利用{0,2,4,5}作为打鼾的赋值,拟合 logistic 回归模型。通过拟合的概率、线性近似,以及对发生比的效应,解释结果。分析模型的拟合优度。

5.4 Hastie 和 Tibshirani(1990,p. 282)描述了一项有关导致驼背的风险因素的研究,即在脊柱矫正手术后出现严重脊柱前倾的情况。18 个出现驼背的对象在手术时的年龄以月为单位分别为 12,15,42,52,59,73,82,91,96,105,114,120,121,130,139,139,157;22 个未出现驼背的对象分别为 1,1,2,8,11,18,22,31,37,61,72,81,97,112,118,127,131,140,151,159,177,206。

a. 以年龄作为预测变量,拟合预测是否出现驼背的 logistic 回归模型。检验年龄的效应是否显著。

b. 将数据进行绘图。指出在是否存在驼背的不同结果中,年龄的离散程度的差异。拟合模型 $\text{Logit}[\pi(x)]=\alpha+\beta_1 x+\beta_2 x^2$。检验年龄的平方项是否显著,将拟合值绘图并加以解释(另见习题 5.33)。

5.5 参见表 6.11。皮尔逊独立性检验结果为 $X^2(I)=6.88(P=0.14)$。利用等距的赋值,Cochran-Armitage 趋势检验结果为 $z^2=6.67(P=0.01)$。解释并说明为什么两个检验的结果如此不同。通过线性 Logit 模型分析该数据。利用沃尔德和似然比方法进行独立性检验,并与 Cochran-Armitage 检验的结果进行比较。检查模型的拟合情况,并加以解释。

5.6 对于表 5.3,利用(1,2,3,4,5)而不是(0.0,0.5,1.5,4.0,7.0)作为对饮酒量的赋值进行趋势检验。比较两组检验的结果,注意分布极不均匀的数据对所选取赋值的敏感性。

5.7 参见表 2.11,使用(0,3,9.5,19.5,37,55)作为对吸烟量的赋值,通过 Logit 模型进行数据分析。对截距的估计有意义吗?加以说明。

5.8 一项研究利用 1998 年风险行为因素社会调查(1998 Behavioral Risk Factors Social Survey)数据考察影响美国妇女服用口服避孕药的相关因素。表 5.13 给出了关于

服用口服避孕药的概率的logisitic回归模型结果。每个预测变量都由虚拟变量来表示,该表列出了当虚拟变量取值为1时的类别。解释这些效应。构建关于服用避孕药和教育水平的条件发生比之比的置信区间并加以解释。

表5.13 习题5.8的数据

变 量	以下情况编码为1	估计值	标准误
年龄	35岁及以下	-1.320	0.087
种族	白人	0.622	0.098
教育程度	1年大学及以上	0.501	0.077
婚姻状况	已婚	-0.460	0.073

来源:由佛罗里达大学药学院的Debbie Wilson友情提供。

5.9 参见表2.6。表5.14给出了Logit模型的拟合结果,将死刑判决视为结果变量(1=是),被告人种族(1=白人)和受害人种族(1=白人)对应的虚拟变量作为预测变量。

表5.14 习题5.9的输出结果

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	1	0.379 8
Pearson Chi-Square	1	0.197 8
Log Likelihood		-209.478 3

Parameter	Estimate	Standard Error	Likelihood Ratio 95% Conf Limits		Chi-Square
Intercept	-3.596 1	0.506 9	-4.775 4	-2.734 9	50.33
def	-0.867 8	0.367 1	-1.563 3	-0.114 0	5.59
vic	2.404 4	0.600 6	1.306 8	3.717 5	16.03

LR Statistics

Source	DF	Chi-Square	Pr > ChiSq
def	1	5.01	0.025 1
vic	1	20.35	<.000 1

译者注——def:被告人种族;vic:受害人种族。

a. 解释对参数的估计。哪一组人最可能被判死刑?求此种情况下的概率估计。

b. 解释关于条件发生比之比的95%的置信区间。

c. 在控制受害人种族的情况下,对被告人种族的效应进行检验:(i)使用沃尔德检验,(ii)使用似然比检验。对结果加以解释。

d. 检验模型的拟合优度,并对结果加以解释。

5.10 对于表2.13,通过模型考察受害人种族和被告人种族的效应。对结果加以解释。

5.11 表5.15取自一项有关15—16岁青少年的全国性调查。这里研究的核心问题为是否曾经有过性行为。分析该数据,描述性别和种族的效应,并进行统计推断;检验模型的拟合优度,并简要地解释结果。

5.12 按照《独立报》的报道(*Independent*, London, Mar. 8, 1994),截至1993年3月伦敦市警方共报告了30 475人失踪,其中年龄在13岁及以下的3 271个失踪的男孩以及2 486个失踪的女孩中,分别有33个和38个在一年后仍然没找到。在14—18岁

之间的7 256个失踪的男孩和8 877个失踪的女孩中,仍未找到的人数分别为63和108;对于19岁及以上的5 065个男性和3 520个女性失踪者,未找到的人数分别为157和159。分析以上数据,并解释结果(感谢Pat Altham提供了上述数据)。

表5.15 习题5.11的数据

种族	性别	性行为	
		是	否
白人	男性	43	134
	女性	26	149
黑人	男性	29	23
	女性	22	36

来源:S. P. Morgan and J. D. Teachman, *J. Marriage Fam.* 50:929-936(1988)。授权重印自:the National Council on Family Relations。

5.13 美国大学体育协会(National Collegiate Athletic Association, NCAA)研究了在1984—1985学年入学的体育生的毕业率。(样本规模,毕业人数)的分布如下:对于白人女生为(796,498),白人男生为(1 625,878),黑人女生为(143,54),以及黑人男生为(197,60)(J. J. McArdle and F. Hamagami, *J. Amer. Statist. Assoc.* 89:1107-1123, 1994)。分析这些数据,并解释结果。

5.14 在一项旨在评估一个教育项目是否会提高有性行为的青少年获取避孕套的可能性的研究中,青少年被随机分为两个实验组。其中一组参加了该教育项目,包括关于艾滋病病毒传播的讲座和录像,而另一组没有参加。表5.16给出了关于青少年获取避孕套的影响因素的logistic模型结果。

a. 前三个预测变量是编码为(1,0)虚拟变量,找出拟合模型的参数估计。利用对数发生比之比的相应置信区间,确定组别效应的标准误。

b. 解释为什么对性别的发生比之比的估计值1.38或者相应的置信区间是错误的。证明如果给出的区间是正确的,那么1.38实际上是对数发生比之比,所估计的发生比之比应为3.98。

表5.16 习题5.14的数据

变量	发生比之比	95%的置信区间
组别(参与教育项目,未参与)	4.04	(1.17,13.9)
性别(男性,女性)	1.38	(1.23,12.88)
社会经济地位(高,低)	5.8	(1.87,18.28)
已有性伴侣的总数	3.22	(1.08,11.31)

来源:V. I. Rickert et al., *Clin. Pediatr.* 31:205-210(1992)。

5.15 表5.17所示为一个以鳞状细胞食道癌($Y=1$,是;$Y=0$,否)为结果变量的logistic回归模型的估计效应。其中对于每天至少抽一盒烟的吸烟者吸烟状况(S)等于1,其他为0;饮酒(A)的取值等于每天的平均饮酒量;种族(R)以黑人等于1,白人为0。为了描述种族与吸烟的交互效应,分别构建当$R=1$和$R=0$时的预测方程,求在每种情况下拟合的YS的条件发生比之比。类似地,构建当$S=1$和

$S=0$ 时的预测方程。求在每种情况下拟合的 YR 的条件发生比之比。注意对于每一种关联,交互乘积项的系数等于在另一个变量的两种给定取值下的对数发生比之比的差。解释为什么 S 的系数代表了白人中 Y 和 S 的对数发生比之比。R 和 S 的 P 值对应的假设是什么?

表 5.17 习题 5.15 的数据

变 量	效 应	P 值
截距	-7.00	<0.01
饮酒	0.10	0.03
吸烟	1.20	<0.01
种族	0.30	0.02
种族×吸烟	0.20	0.04

5.16 在一项关于高中生的调查中,Y = 被调查者是否曾经在过量饮酒后驾车(1 = 是) s = 性别(1 = 女性),r = 种族(1 = 黑人;0 = 白人),g = 年级($g_1=1$,9 年级;$g_2=1$ 10 年级;$g_3=1$,11 年级;$g_1=g_2=g_3=0$,12 年级)。预测方程为

$$\text{Logit}[\hat{P}(Y=1)] = -0.88-0.40s-0.72r-2.22g_1-1.43g_2-0.58g_3+0.74rg_1+0.38rg_2+0.01rg_3。$$

a. 详细解释这些效应。通过描述在每一年级的种族效应以及每一种族的年级效应来解释交互项。

b. 用 r_1(1 = 黑人,0 = 其他)替代上面的 r,另外包括 r_2(1 = 西班牙裔,0 = 其他),并且当 $r_1=r_2=0$ 时对应为白人。假定预测方程和上面相同,但是多了一些附加项 $-0.29r_2+0.53r_2g_1+0.25r_2g_2-0.06r_2g_3$。解释这些效应。

5.17 表 5.18 给出了关于 Y = 进行全身麻醉手术的病人在醒来时是否会感到喉咙痛(0 = 否,1 = 是)的研究结果,其中预测变量为 D = 手术用时(以分钟为单位)和 T = 导气管种类(0 = 喉罩气道,1 = 气管导管)。利用这些预测变量拟合 Logit 模型,解释关于参数的估计,并进行统计推断。

表 5.18 习题 5.17 的数据

病人	D	T	Y	病人	D	T	Y	病人	D	T	Y
1	45	0	0	13	50	1	0	25	20	1	0
2	15	0	0	14	75	1	1	26	45	0	1
3	40	0	1	15	30	0	0	27	15	1	0
4	83	1	1	16	25	0	1	28	25	0	1
5	90	1	1	17	20	1	0	29	15	1	0
6	25	1	1	18	60	1	1	30	30	0	1
7	35	0	1	19	70	1	1	31	40	0	1
8	65	0	1	20	30	0	1	32	15	1	0
9	95	0	1	21	60	0	1	33	135	1	1
10	35	0	1	22	61	0	0	34	20	1	0
11	75	0	1	23	65	0	1	35	40	1	0
12	45	1	1	24	15	1	0				

来源:数据取自:D. Collett, in *Encyclopedia of Biostatistics*(New York:Wiley:1998),*PP*. 350-358。

5.18 参考利用 x = 壳宽来分析马蹄蟹数据的模型(式 5.2)。

a. 给出(i)在壳宽取均值 26.3 时,估计的拥有同伴的发生比等于 2.07;(ii)当 x = 27.3 时,估计的发生比等于 3.40;(iii)由于 $\exp(\hat{\beta}) = 1.64$, $3.40 = 1.64 \times 2.07$,因此发生比增加了 64%。

b. 基于 β 的 95% 的置信区间,证明在 $\pi = 0.5$ 附近时,x 每增加 1 厘米所对应的拥有同伴的概率的增长率在 0.07 和 0.17 之间。

5.19 对于表 4.3,利用颜色作为唯一的预测变量,拟合关于拥有同伴的概率的 logistic 回归模型。

a. 将颜色作为定类变量,解释为什么这个模型是饱和的。通过每种颜色的样本 Logit 表达模型的参数估计。

b. 构建关于颜色效应为零的似然比检验。

c. 将颜色视为定量变量重新拟合模型。解释模型的拟合结果,并对颜色效应为零进行检验。

d. 检验(c)部分的模型拟合优度。加以解释。

5.20 参考式 5.14 模型。分别对 $c = 1$ 和 $c = 0$ 的情况,求在壳宽的第一个和第三个四分位点拥有同伴的估计概率,以描述壳宽的效应。

5.21 参考模型(5.13)的预测方程 $\text{Logit}(\hat{\pi}) = -10.071 - 0.509c + 0.458x$。颜色的均值和标准差分别为 $\bar{c} = 2.44$ 和 $s = 0.80$,壳宽的分别为 $\bar{x} = 26.30$ 和 $s = 2.11$。对于标准化的预测变量(例如,x = (壳宽 − 26.3)/2.11),说明为什么 c 和 x 的估计系数等于 −0.41 和 0.97。通过比较每个预测变量每增加 1 个标准差对发生比的偏效应来解释上述结果。估计当壳宽取均值时在第一种和最后一种颜色之间 $\hat{\pi}$ 的变化,并以此描述颜色的效应。

5.22 参考式 5.12 模型。

a. 利用 x = 体重拟合模型。解释体重和颜色的效应。

b. 允许交互项的模型是否拟合得更好? 加以解释。

c. 对于(b)部分,构建一个关于浅褐色和黑色蟹的斜率参数之差的置信区间,加以解释。

d. 将颜色视为定量变量,重做(a)到(c)部分。

5.23 Fowlkes 等(1988)报告了一项针对全国性大企业员工的调查结果,以研究工作满意度与种族、性别、年龄以及地区的关系。该数据可从本书的网站上(*www.stat.ufl.edu/~aa/cda/cda.html*)找到。对这些数据拟合 Logit 模型并解释参数估计结果。Fowlkes 等(1988)报告说:"最不满的雇员是那些年龄低于 35 岁、女性、其他(种族),以及在东北部工作的雇员;……最满意的为年龄大于 44 岁、男性、其他,以及在太平洋或中西部地区工作的雇员;这样的雇员感到满意的发生比大约为 3.5比 1。"解释如何通过模型拟合结果得出上述结论。

5.24 令 Y 表示调查对象对现行法律中允许人工流产的态度(1 = 支持),h 为性别(h = 1,女性;h = 2,男性),i 为宗教(i = 1,新教;i = 2,天主教;i = 3,犹太教),j 为政党(j = 1,民主党;j = 2,共和党;j = 3,独立党派)。利用调查数据,统计软件对模型 $\text{Logit}[P(Y=1)] = \alpha + \beta_h^G + \beta_i^R + \beta_j^P$ 的拟合结果为 $\hat{\alpha} = 0.62$, $\hat{\beta}_1^G = 0.08$, $\hat{\beta}_2^G = -0.08$, $\hat{\beta}_1^R = -0.16$, $\hat{\beta}_2^R = -0.25$, $\hat{\beta}_3^R = 0.41$, $\hat{\beta}_1^P = 0.87$, $\hat{\beta}_2^P = -1.27$, $\hat{\beta}_3^P = 0.40$。

a. 解释宗教如何影响支持态度的发生比。

b. 对于最有可能和最不可能支持现行法律的组别,分别估计其支持概率。

c. 如果对参数的限定条件为 $\beta_1^G=\beta_1^R=\beta_1^P=0$,给出估计结果。

5.25 表5.19所示为一项关于收入与是否拥有旅行信用卡之间关系的意大利研究所随机选出的样本。该表给出了每个年收入水平(以百万里拉为单位)对应的样本数量以及至少拥有一张旅行信用卡的人数。分析这些数据。

5.26 参见表9.1,将吸食大麻作为结果变量,分析这些数据。

5.27 本书的网站(*www.stat.ufl.edu/~aa/cda/cda.html*)包括一个关于职业抱负(高、低)与性别、居住地、智商以及社会经济地位的关系的五维表格。分析这些数据。

理论与方法

5.28 对于式5.1模型,证明 $\partial\pi(x)/\partial x=\beta\pi(x)[1-\pi(x)]$。

5.29 对于式5.1模型,当 $\pi(x)$ 很小时,说明为什么可以将 $\exp(\beta)$ 近似解释为 $\pi(x+1)/\pi(x)$。

5.30 证明 logistic 回归曲线(式5.1)在 $\pi(x)=\frac{1}{2}$ 时具有最陡的斜率。将结果扩展到式5.8模型。

表5.19 习题5.25的数据

收入/百万里拉	研究对象的数量	拥有信用卡的人数	收入/百万里拉	研究对象的数量	拥有信用卡的人数	收入/百万里拉	研究对象的数量	拥有信用卡的人数
24	1	0	39	2	0	65	6	6
27	1	0	40	5	0	68	3	3
28	5	2	41	2	0	70	5	3
29	3	0	42	2	0	79	1	0
30	9	1	45	1	1	80	1	0
31	5	1	48	1	0	84	1	0
32	8	0	49	1	0	94	1	0
33	1	0	50	10	2	120	6	6
34	7	1	52	1	0	130	1	1
35	1	1	59	1	0			
38	3	1	60	5	2			

来源:*Categorical Data Analysis*, Quaderni del Corso Estivo di Statistica e Calcolo delle Probabilità, n. 4., Istituto di Metodi Quantitativi, Università Luigi Bocconi, by R. Piccarreta。

5.31 校准(calibration)问题就是要估计满足 $\pi(x)=\pi_0$ 的 x 的取值。对于线性 Logit 模型,论证它的置信区间是一组满足以下条件的 x 值。

$$|\hat{\alpha}+\hat{\beta}x-\text{Logit}(\pi_0)|/[\text{var}(\hat{\alpha})+x^2\text{var}(\hat{\beta})+2x\text{cov}(\hat{\alpha},\hat{\beta})]^{1/2}<z_{\alpha/2}。$$

Morgan(1992, Sec. 2.7)综述了其他方法。

5.32 一项研究考察职业体育中从某一年有潜力的运动员中进行选秀的情况。该研究分析一个运动员在选秀中的位置 $d(d=1,2,3,\cdots)$ 对其最终成为全明星的概率 π 的影响,所使用的模型为 $\text{Logit}(\pi)=\alpha+\beta\log d$(S. M. Berry, *Chance*, 14:53-57, 2001)。

a. 证明 $\pi/(1-\pi)=e^{\alpha}d^{\beta}$,并证明 e^{α} = 状元秀的发生比。

b. Berry 报告说，在美国，对于职业篮球 $\hat{\alpha}=2.3$，$\hat{\beta}=-1.1$；对于职业棒球，$\hat{\alpha}=0.7$，$\hat{\beta}=-0.6$。这表明，在篮球中状元签更关键，其他的高签位成为全明星的可能性相对较小。说明为什么。

5.33 对于具有 $Y=j$ 的研究对象的总体，X 服从 $N(\mu_j,\sigma^2)$ 分布，$j=0,1$。

a. 利用贝叶斯定理，证明 $P(Y=1|x)$ 满足 $\beta=(\mu_1-\mu_0)/\sigma^2$ 的 logistic 回归模型。

b. 假定 $(X|Y=j)$ 服从 $N(\mu_j,\sigma_j^2)$，其中 $\sigma_0\neq\sigma_1$，证明 logistic 模型在包括一个二次项时成立(Anderson,1975)(习题 5.4 表明，当在 $y=0$ 和 $y=1$ 的情况下所对应的 x 值的离散度差别很大时，有必要在模型中包括一个二次项。这个结果也意味着，当方差存在差别时，为了检验正态分布均值是否相等，可以拟合一个以两个分组为结果变量的带有二次项的 logistic 回归，并对二次项进行检验；参见：O' Brein(1988)。

c. 假定 $(X|Y=j)$ 服从指数离散族分布，密度函数为 $f(x;\theta_j)=\exp\{[x\theta_j-b(\theta_j)]/a(\phi)+c(x,\phi)\}$。求出相应的 logistic 模型。

d. 对于多个预测变量，假定 $(\mathbf{X}|Y=j)$ 服从多元分布 $N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma})$，$j=0,1$。证明 $P(Y=1|\mathbf{x})$ 满足效应参数为 $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)$ 的 logistic 回归(Cornfield,1962)。

5.34 对于某一严格递增的累积分布函数 F，假定 $\pi(x)=F(x)$，说明为什么会存在关于 x 的单调转换使得 logistic 回归模型成立。将结果扩展到其他连结函数的情形。

5.35 对于 $I\times 2$ 表格，考虑 Logit 式 5.4 模型。

a. 给定 $\{\pi_i>0\}$，指出如何得到满足 $\beta_I=0$ 的 $\{\beta_i\}$。

b. 证明 $\beta_1=\beta_2=\cdots=\beta_I$ 表示独立性模型。给出相应的似然方程，并证明 $\hat{\alpha}=\text{Logit}[(\sum_i y_i)/(\sum_i n_i)]$。

5.36 一项研究存在独立的二项分布结果，在 $x=0$ 的 n_0 次试验中成功次数为 y_0，在 $x=1$ 的 n_1 次试验中成功次数为 y_1。构建模型 $\text{Logit}[\pi(x)]=\alpha+\beta x$ 的对数似然函数。推导它的似然方程，并证明 $\hat{\beta}$ 等于样本的对数发生比之比。

5.37 一项研究在 $X=x_i$ 时存在 n_i 个独立的二项分布观测值 $\{y_{i1},\cdots,y_{in_i}\}$，$i=1,\cdots,N$，并且 $n=\sum_i n_i$。考虑模型 $\text{Logit}(\pi_i)=\alpha+\beta x_i$，其中 $\pi_i=P(Y_{ij}=1)$。

a. 证明无论将数据视为 n 个伯努利观测值还是 N 个二项分布观测值，似然函数的核函数是相同的。

b. 对于饱和模型，说明为什么在这两种数据形式下似然函数不一样(提示：参数的数目不同)。因此，软件所报告的偏离度取决于数据存储的形式。

c. 说明为什么两个非饱和模型的偏离度之差不取决于数据存储的形式。

d. 假定所有的 $n_i=1$，证明偏离度取决于 $\hat{\pi}_i$ 而不取决于 y_i。因此，检查模型的拟合优度是没有意义的(另见习题 4.22)。

5.38 假定 Y 服从 $\text{bin}(n,\pi)$ 分布。对于模型 $\text{Logit}(\pi)=\alpha$，检验 H_0：$\alpha=0$(即，$\pi=0.5$)。令 $\hat{\pi}=y/n$。

a. 由第 3.1.6 节可知，$\hat{\alpha}=\text{Logit}(\hat{\pi})$ 的渐近方差等于 $[n\pi(1-\pi)]^{-1}$。利用检验统计量 $[\text{Logit}(\hat{\pi})/\text{SE}]^2$，比较沃尔德检验的估计标准误(SE)与使用 π 的初始值所估计的标准误。证明沃尔德统计量与使用初始 SE 的统计量的比等于 $4\hat{\pi}(1-\hat{\pi})$。如果 $|\alpha|$ 很大并且 $\hat{\pi}$ 的值接近 0 或 1，这对沃尔德检验的性能意味着什么？

b. 沃尔德推断取决于参数设定方式。在 $[(\hat{\pi}-0.5)/\text{SE}]^2$ 下，上述的统计量之比

检验的如何变化？其中 SE 是关于 $\hat{\pi}$ 的估计 SE 或者初始 SE。

c. 假定 $y=0$ 或 $y=n$，证明在(a)部分的沃尔德检验中，对于任意 $0<\pi_0<1$ 都无法拒绝 $H_0:\pi=\pi_0$，而(b)部分的沃尔德检验则会拒绝掉所有这样的 π_0。(注意：相似的结果也适用于有关泊松分布均值与对数均值的推断；参见：Mantel (1987a))。

5.39 给出式 5.10 模型的似然方程。证明这些方程意味着，在边际二维表中，拟合值与样本值相等。

5.40 考虑关于 $I\times 2$ 表格的线性 Logit 式 5.5 模型，其中 y_i 是服从 $\text{bin}(n_i,\pi_i)$ 的变量。

a. 证明模型的对数似然函数为

$$L(\beta)=\sum_{i=1}^{I}y_i(\alpha+\beta x_i)-\sum_{i=1}^{I}n_i\log[1+\exp(\alpha+\beta x_i)]。$$

b. 证明 β 的充分统计量为 $\sum_i y_i x_i$，并说明为什么这实际就是在 Cochran-Armitage 检验中所使用的变量(因而该检验是关于 $H_0:\beta=0$ 的计分检验)。

c. 令 $S=\sum_i y_i$，证明似然方程为

$$S=\sum_i n_i\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}$$

$$\sum_i y_i x_i=\sum_i n_i x_i\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}。$$

d. 令 $\{\hat{\mu}_i=n_i\hat{\pi}_i\}$，说明为什么 $\sum_i\hat{\mu}_i=\sum_i y_i$，并且

$$\sum_i x_i\frac{y_i}{S}=\sum_i x_i\frac{\hat{\mu}_i}{\sum_a\hat{\mu}_a}。$$

解释为什么这意味着模型拟合的第一列中所有行关于 x 的平均赋值与所观察到的数据相同。在第二列中也是如此。

5.41 令 Y_i 在 x_i 处服从 $\text{bin}(n_i,\pi_i)$，并令 $p_i=y_i/n_i$。对于使用 Logit 连结的二项分布广义线性模型：

a. 对于邻近 π_i 的 p_i，证明

$$\log\frac{p_i}{1-p_i}\approx\log\frac{\pi_i}{1-\pi_i}+\frac{p_i-\pi_i}{\pi_i(1-\pi_i)}。$$

b. 证明式 5.23 中的 $z_i^{(t)}$ 是第 i 个样本 Logit 的线性形式在 $\hat{\pi}_i$ 的近似值 $\pi_i^{(t)}$ 处的取值。

c. 验证关于 $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$ 的公式(式 5.20)。

5.42 通过图表说明在以下情况下，利用解释变量 X 和 Z 对结果变量 Y 进行模型分析时不存在交互效应(*no interaction*)分别指什么意思：

a. 所有的变量都是连续变量(多元回归)。

b. Y 和 X 是连续变量，Z 是分类变量(协方差分析)。

c. Y 是连续变量，X 和 Z 是分类变量(二维方差分析)。

d. Y 是二分变量，X 和 Z 是分类变量(Logit 模型)。

6 Logistic 回归模型的构建与应用

在介绍了 logistic 回归模型的拟合和解释等基础知识后,我们现在讨论 logistic 回归模型的构建与应用。当存在多个解释变量时,可能会有许多可供选择的模型。在第 6.1 节,我们讨论模型选择的策略。在选定一个初步的模型后,模型检查考虑该模型是否存在系统性的拟合不足。第 6.2 节的内容介绍模型检查的诊断方法,如残差分析。

在现实中,一种常见的应用是,将数据按照控制变量分层,比较二分结果变量在两组间的差异。在第 6.3 节,我们介绍对这种数据的 Logit 分析。第 6.4 节讨论精心选取的模型在对关联进行检测和估计方面提高推断效能的优势。第 6.5 节的内容为 logistic 回归的统计效能以及样本规模的决定因素。尽管对于概率而言,Logit 是最常用的连结函数,但在特定情况下可能其他的连结函数更为适用。在第 6.6 节中,我们介绍应用 probit 连结以及双对数连结的模型。

在小样本或者模型参数很多的情况下,普通的大样本最大似然推断可能并不适用。在第 6.7 节中,我们讨论*条件 logistic 回归*(*conditional logistic regression*)。与 2 × 2 表格的小样本方法相同,该模型通过条件化方法来消除冗余参数的影响。

6.1 模型选择的策略

Logistic 回归模型的选择与普通回归所面临的问题是一样的。随着解释变量数量的增加,可能的效应和交互项会迅速增多,因而选择适当模型的过程变得更加困难。这里存在两个竞争性目标:一方面,模型应当足够复杂以保证能较好地拟合数据;另一方面,模型应当便于解释,即它是对数据的修匀而不能过度拟合。

多数研究都是为了回答某些特定的问题,这些问题引导着对模型所包含项的选择。在验证性分析(confirmatory analysis)中,我们通常会比较一组模型。例如,针对某种效应的研究假设,可以比较包括和不包括该效应的模型来进行检验。对于探索性研究,在可能成立的模型中进行选择,可以提供有关变量间关联结构的信息,并提出在未来研究中需要关注的问题。

无论哪一种情况,首先考察每个预测变量分别对 Y 的效应总是有益的。如果预测变量是连续的,可以通过绘图的方式(与修匀相结合)进行研究;如果是离散的,则通过列联表的方式。这种分析可以对边际效应提供一种大致的认识。不均衡的数据,即结果变量落在某一类别的观测值数量偏少,会限制模型中预测变量的数目。一个指导原则是,对于每个预测变量,结果变量的每个类别都至少包括 10 个观测值(Peduzzi et al. ,1996)。

例如,如果在 $n=1\ 000$ 中 $y=1$ 只出现了 30 次,模型应当包括不超过三个 x 项。这种指导原则只是大概的,而且它并不意味着在每种类别的结果都有 500 个观测值的情况下,包括 50 个预测项的模型就会表现得很好。

现实中存在多种模型选择方法,但是没有一种在任何情况下都是最好的。在普通回归中需要注意的问题,对于所有广义线性模型仍然适用。例如,包含多个预测变量的模型可能会存在*多重共线性*(*multicollinearity*)问题——即由于预测变量间的相关关系,当所有预测变量都放入模型后,这些变量都不显著。因为当某一变量与模型中的其他变量存在明显的重合时,它的效应就会显得很小。这时,剔除掉多余的预测变量很重要,这样可以降低其他效应估计的标准误。

6.1.1 例子:再析马蹄蟹数据

表 4.3 中的马蹄蟹数据包括四个预测变量:颜色(四个类别),蟹刺状况(三个类别),体重,以及壳宽。现在,我们利用所有这些预测变量来拟合雌蟹是否拥有同伴($y=1$)的 logistic 回归模型。

首先,我们拟合仅包括主效应的模型作为分析的起点,

$$\text{Logit}[P(Y=1)] = \alpha + \beta_1\text{weight} + \beta_2\text{width} + \beta_3 c_1 + \beta_4 c_2 + \beta_5 c_3 + \beta_6 s_1 + \beta_7 s_2,$$

将颜色(c_i)和蟹刺状况(s_j)作为定类变量,通过虚拟变量表示前三种颜色和前两种蟹刺状况。表 6.1 给出了模型结果。关于 Y 联合独立于所有这些预测变量的似然比检验对应的假设为 $H_0:\beta_1=\cdots=\beta_7=0$。检验统计量等于 40.6,自由度为 df = 7($P<0.000\ 1$)。检验结果表明,至少有一个预测变量具有显著效应。

表 6.1 对马蹄蟹数据所拟合的包括所有主效应的模型输出结果

Testing Global Null Hypothesis: BETA = 0

Test	Chi-square	DF	Pr > Chisq
Likelihood Ratio	40.556 5	7	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Std Error	Chi-Square	Pr > Chisq
截距	-9.273 4	3.837 8	5.838 6	0.015 7
体重	0.825 8	0.703 8	1.376 5	0.240 7
壳宽	0.263 1	0.195 3	1.815 2	0.177 9
第一种颜色	1.608 7	0.935 5	2.956 7	0.085 5
第二种颜色	1.505 8	0.566 7	7.060 7	0.007 9
第三种颜色	1.119 8	0.593 3	3.562 4	0.059 1
第一种蟹刺状况	-0.400 3	0.502 7	0.634 0	0.425 9
第二种蟹刺状况	-0.496 3	0.629 2	0.622 2	0.430 2

尽管模型总体检验极为显著,表 6.1 的结果却并不理想。模型中关于体重和壳宽的参数估计仅略大于相应的标准误。定类变量的估计将各组与参照组(最后一组)进行比较。就颜色而言,最大的组间差别也不足标准误的两倍;就蟹刺状况而言,各组间最大的差别都小于一个标准误。

总体检验所对应的 P 值很小,不过个体效应却不显著,这是存在多重共线性问题的一个信号。在第 5.2.2 节中,我们发现了存在着很强的壳宽效应。当控制了体重、颜色,以及蟹刺状况后,壳宽的偏效应不再显著。然而,体重和壳宽之间存在强相关(0.887)。从实用的角度来说,这两个变量都是很好的预测变量,但是将二者都包括进来却是冗余的。在接下来的分析中,我们使用壳宽(W)、颜色(C)和蟹刺状况(S)作为预测变量。简便起见,我们利用每个变量的最高阶项来表示模型,将 C 和 S 视为定类变量。例如,($C+S+W$)代表一个只包括主效应的模型,而($C+S*W$)则代表一个包括所有主效应并加上 $S*W$ 的交互项的模型。一般来说,只考虑交互效应而不考虑组成交互项的变量的主效应是不合理的。

6.1.2 逐步建模法

在探索性研究中,如果我们能够审慎地使用结果,可以考虑使用一种算法在多个模型中寻找最优模型。Goodman(1971a)提出了与普通回归中的前向选择(forward selection)和后向剔除(backward elimination)相类似的方法。

前向选择法将每一项逐步加入到回归中,直到模型的拟合程度不再随着新加入的回归项而提高为止。在每一步,它都优先挑选能够最大程度改进模型拟合的项。由于不同项所导致的偏离度的消减可能对应着不同的自由度,对模型中各项进行检验的最小 P 值是一个合理的选择标准。这一过程在每一步都重新检验前面已加入的项是否仍然显著。

后向剔除法以一个复杂模型为起点,然后逐步剔除贡献较小的各项。在每一步,它选出去除后对模型拟合影响最小的一项(即,具有最大的 P 值)。当再剔除任何一项都会显著地影响模型的拟合时,我们就得到了最终的模型。就具有多于两个类别的定类变量来说,对于每一种方法,程序在每一步都应该考虑整个变量而不是单个的虚拟变量。也即,加入或者剔除整个变量而不是它的某一个虚拟变量,否则,回归结果会依赖于变量编码本身。同样的道理也适用于包含该变量的交互项。

许多统计学家在两种模型选择方法中更偏好后向剔除法,他们认为从一个过于复杂的模型中删除项比在一个过于简单的模型添加项更安全。前向选择可能会由于序列中的某一具体的检验效能低下而过早中止。然而,两种方法都不必然保证会给出一个有意义的模型。对这些变量选择方法的使用一定要慎重!当你在评估许多项时,一两个并不重要的项可能会由于偶然的因素而显得很重要。例如,当所有的真实效应都比较弱时,最大的样本效应可能会导致对真实效应的过度高估。通过多重检验来调整 P 值的办法,可参见:Westfall and Wolfinger(1997)、Westfall and Young(1993)。

一些软件还提供其他的模型选择方法。其中一种方法是,按照某一标准,在预测项数目固定的情况下确定最优模型。如果这种方法得到的结果与后向或前向选择法所得模型相去甚远,则这些模型非常值得怀疑。另外,当同样的方法用于相同规模的随机再抽样样本(bootstrap sample)所得到的模型不一致时,也意味着模型结果存在问题。

最后,统计显著性不应当作为在模型中是否包含某一项的唯一标准。即使在统计上不显著,将一个对研究目的本身至关重要的变量包括进来并报告它的估计效应是合理的。将其包括在模型中,可能有助于降低对其他预测变量效应的估计偏差,并使与其他存在显著效应的研究(可能由于样本规模更大)进行对比成为可能。自动选择模型的程序绝不能替代在模型构建过程中的细致思考。

6.1.3 后向剔除法:以马蹄蟹数据为例

表6.2给出了利用预测变量壳宽、颜色以及蟹刺状况对马蹄蟹数据拟合的几个 Logit 模型的结果。拟合结果中的偏离度检验(G^2)将相应的模型与饱和模型进行了对比。正如在第5.2.4和5.2.5节所指出的,当预测变量为连续变量时,如壳宽,该检验并不近似于卡方分布。但是,参数数量相差较小的两个模型之间的偏离度之差仍然是有意义的。这个差值是比较模型的似然比统计量 $-2(L_0 - L_1)$,它近似服从卡方零分布。

表6.2 关于马蹄蟹数据的几个 logistic 回归模型的拟合结果

	模型的预测变量[a]	偏离度 G^2	df	AIC	模型对比	偏离度之差	相关系数 $r(y,\hat{\mu})$
1	$(C*S*W)$	170.44	152	212.4	—	—	
2	$(C*S+C*W+S*W)$	173.68	155	209.7	(2) - (1)	3.2(df=3)	
3a	$(C*S+S*W)$	177.34	158	207.3	(3a) - (2)	3.7(df=3)	
3b	$(C*W+S*W)$	181.56	161	205.6	(3b) - (2)	7.9(df=6)	
3c	$(C*S+C*W)$	173.69	157	205.7	(3c) - (2)	0.0(df=2)	
4a	$(S+C*W)$	181.64	163	201.6	(4a) - (3c)	8.0(df=6)	
4b	$(W+C*S)$	177.61	160	203.6	(4b) - (3c)	3.9(df=3)	
5	$(C+S+W)$	186.61	166	200.6	(5) - (4b)	9.0(df=6)	
6a	$(C+S)$	208.83	167	220.8	(6a) - (5)	22.2(df=1)	
6b	$(S+W)$	194.42	169	202.4	(6b) - (5)	7.8(df=3)	
6c	$(C+W)$	187.46	168	197.5	(6c) - (5)	0.8(df=2)	0.452
7a	(C)	212.06	169	220.1	(7a) - (6c)	24.5(df=1)	0.285
7b	(W)	194.45	171	198.5	(7b) - (6c)	7.0(df=3)	0.402
8	$(C=\text{dark}+W)$	187.96	170	194.0	(8) - (6c)	0.5(df=2)	0.447
9	None	225.76	172	227.8	(9) - (8)	37.8(df=2)	0.000

a C,颜色;S,蟹刺状况;W,壳宽

我们使用后向剔除法来选取模型。这里,我们仅检验每一个变量的最高阶项。比如,如果模型包括某一变量的交互项,那么剔除该变量的主效应是不适当的。

我们从最复杂的模型开始,即表6.2中的模型1,由($C*S*W$)表示。该模型包括了每个变量的主效应、三个二维的交互项以及一个三维的交互项。它允许在每一个 CS 的取值组合处都具有不同的壳宽效应(事实上,在一些取值组合处,y 仅仅出现了一种类型的结果,因而这些效应是无法估计的)。将该模型与去除三维交互项的较简单模型($C*S+C*W+S*W$)进行比较的似然比统计量等于3.2(df=3)。这表明,三维交互项不是必须的($P=0.36$)。这样,我们继续模型简化的过程。

下一步,考虑去除掉一个二维交互项的三个模型。其中,模型($C*S+C*W$)给出了与更复杂模型相同的拟合,因而可以剔除 $S\times W$ 交互项。接下来,考虑去除另外两个二维交互项中的一个。剔除 $C\times S$ 交互项的模型($S+C*W$)所对应的偏离度增加了8.0,自由度为6($P=0.24$);剔除 $C\times W$ 交互项的模型所对应的偏离度增加了3.9,自由度为3($P=0.27$)。两个偏离度的变化都不显著,这意味着我们可以去除其中任何一项,并继续简化过程。无论哪种情况,去除掉剩下的交互项似乎也是可以的。比如,从模型($W+C*S$)中剔除 $C\times S$ 交互项,得到模型($C+S+W$),相应的偏离度提高了9.0,对应的自由度为6($P=0.17$)。

现在,模型中只包括主效应。下一步,我们考虑剔除某个主效应。表6.2显示,去掉

S 项对模型拟合几乎没有影响。余下的变量（ C 和 W ）都对模型拟合存在重要影响。比如，去掉 C 导致偏离度（对比模型 7b 和 6c）上升 7.0，自由度为 3（ $P = 0.07$ ）。第 5.4.6 节的分析揭示了黑色蟹（第 4 组）与其他颜色存在明显的差别。将颜色表示为单一虚拟变量（黑色为 0，其他颜色为 1）的较简单模型同样拟合得很好（模型 8 和 6c 之间的偏离度的差等于 0.5，自由度为 2）。这时，再对模型进行任何简化都会导致偏离度的显著增加，因而是不可取的。

6.1.4 AIC、模型选择与正确模型

在选取模型时，如果我们认为已经找到了真实的模型，那肯定是错误的。任何模型都是对现实的一种简化。举例来说，无论使用 Logit 连结还是恒等连结，壳宽对拥有同伴的概率的效应都不会是完全线性的。

那么，既然我们知道模型并不真正成立，那对模型拟合进行检验的逻辑是什么？一个拟合充分的简单模型具有简约性（parsimony）的优点。如果模型具有相对很小的偏误，对现实描述得很好，它往往会对所关注的参数给出更准确的估计。我们在第 3.3.7 节和第 5.2.2 节对此进行了讨论，并将在第 6.4.5 节进一步加以探讨。

除了显著性检验，其他标准也可以帮助我们选择一个对所研究的特质进行较好估计的模型。最为大家所熟知的一种标准叫做 *Akaike 信息准则*（*Akaike information criterion*，AIC）。它通过比较模型的拟合值与真值（按照某种期望值）的接近程度来评判一个模型。即便一个简单模型比更复杂的模型偏离真实模型更远，由于它通常会给出关于真实模型的某些特征的更好估计，如单元格概率，我们可能会偏好简单模型。因而，最优模型是能够提供对现实情况最接近的拟合的模型。给定一个样本，Akaike 信息准则会选取使

$$\text{AIC} = -2(\text{最大对数似然值} - \text{模型中参数的数目})$$

最小化的模型。AIC 对包括大量参数的模型进行了某种“惩罚”。就关于分类变量 Y 的模型而言，这种排序等价于一种对偏离度的调整，$[G^2 - 2(\text{df})]$，即减去二倍的残差自由度。有关支持使用 AIC 的有力论证，参见：Burnham and Anderson（1998）。

我们通过表 6.2 所列的模型来演示如何使用 AIC 进行模型选择。该表给出了各模型所对应的 AIC 值。在所有使用三个基本变量的模型中，模型 $C + W$ 所具有的 AIC 值最小（AIC = 197.5），它包括颜色和壳宽的主效应。用颜色是否为黑色的虚拟变量拟合模型，相应更简单的模型表现得更好（AIC = 194.0）。这两个模型看上去都很好。这时，我们应当结合模型 $C + W$ 的拟合结果，综合评判更简单模型对应的较低 AIC 值。

6.1.5 根据因果假设构建模型

尽管模型选择程序是有用的探索性工具，模型构建的过程应当以理论和常识为依据。一般情况下，变量发生的时间顺序反映了可能存在的因果关系。这样，逐次分析一系列模型有助于探究这些关系（Goodman，1973）。

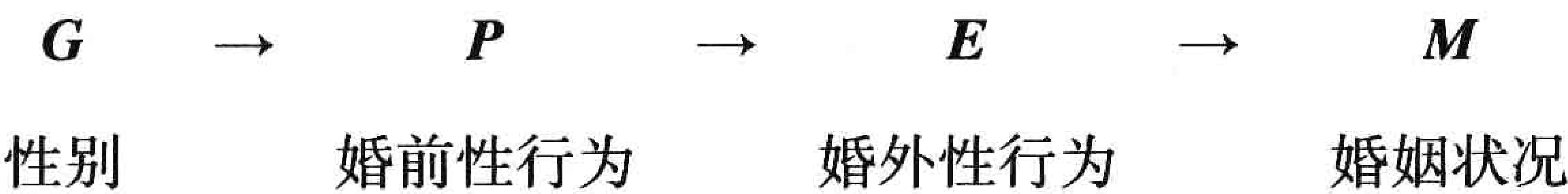
我们以表 6.3 的一项英国研究为例来对此加以说明。调查样本包括申诉过离婚的男性和女性以及数量相当的在婚者，他们被问到：（a）“在你和（过去的）丈夫/妻子结婚前，有没有和他人发生过性行为？”；（b）“在（过去的）婚姻期间，你有没有过婚外情或一夜情？”这个 $2 \times 2 \times 2 \times 2$ 表格包括四个变量：G = 性别，E = 是否有婚外性行为，P = 是否有婚前性行为，M = 婚姻状况。

表 6.3 按照婚前、婚外性行为情况划分的婚姻状况

		性别							
		女性				男性			
	婚前性行为	是		否		是		否	
婚姻状况	婚外性行为	是	否	是	否	是	否	是	否
离异		17	54	36	214	28	60	17	68
仍在婚		4	25	4	322	11	42	4	130

来源：G. N. Gilert, *Modelling Society* (London: George Allen & Unwin, 1981)。授权重印自：Unwin Hyman Ltd.

这四个变量所发生的时点意味着以下的变量顺序：



当位于右边的变量是结果变量时，它前面的变量都是解释变量。图 6.1 给出了一种可能的因果结构。在这个图中，箭头所指的变量在分析的某个阶段就是模型的结果变量。解释变量则会有箭头直接地或间接地指向结果变量。

首先，我们将 P 视为结果变量。图 6.1 预测 G 对 P 存在直接效应，因此假定它们之间相互独立的模型是不充分的。在第二阶段，E 是结果变量。图 6.1 预测 P 和 G 对 E 具有直接效应，同时，G 通过对 P 的影响而对 E 具有间接效应。这些对 E 的效应可以通过包括 G 和 P 的主效应的 Logit 模型来分析。如果 G 对 E 只存在间接效应，则模型中只包括 P 作为预测变量就可以了；也就是说，在控制了 P 后，E 和 G 是条件独立的。在第三阶段，M 是结果变量。图 6.1 预测 E 对 M 有直接效应，P 对 M 既存在直接效应也存在通过 E 的间接效应，同时，G 通过对 P 和 E 的影响对 M 具有间接效应。这意味着，关于 M 的 Logit 模型应当包括 E 和 P 的主效应。对于这个模型而言，在给定 P 和 E 的情况下，G 和 M 是独立的。

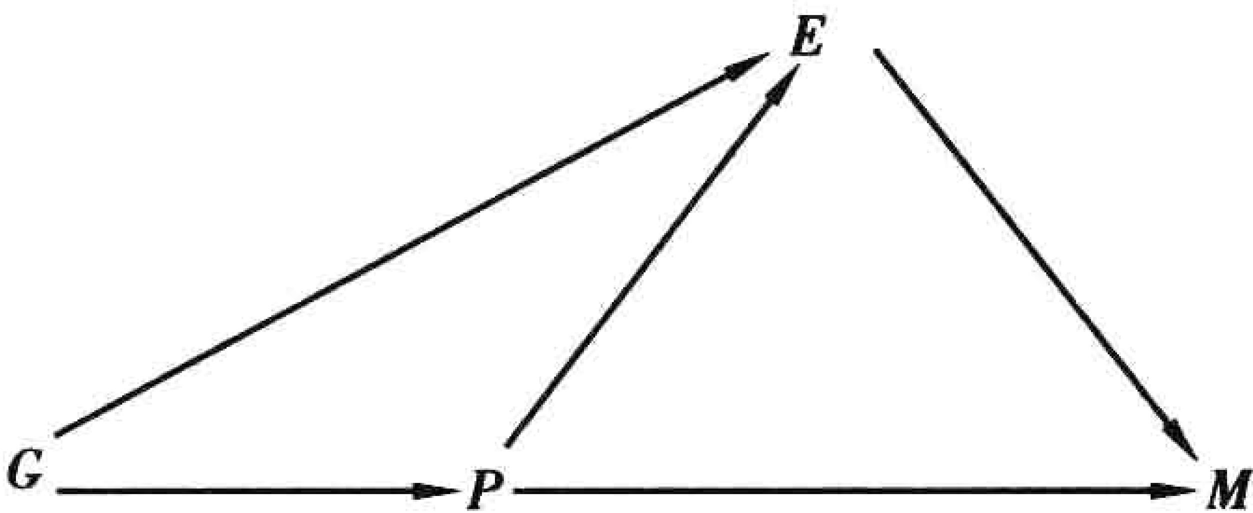


图 6.1 关于表 6.3 的因果关系图

表 6.4 给出了模型分析结果。第一阶段，以 P 作为结果变量，结果显示存在很强的 GP 关联。二者对应的边际表的样本发生比之比为 0.27，也即，所估计的女性发生婚前性行为的发生比是男性的 0.27 倍。第二阶段以 E 作为结果变量。在模型中包括了 P 的效应后，将 G 对 E 的直接效应和间接效应加入模型只得到了微弱的支持，G^2 下降了 2.9 (df = 1)。对于这个模型，EP 的条件发生比之比的估计值为 4.0。

第三阶段以 M 作为结果变量。图 6.1 指定了包括 E 和 P 的主效应的 Logit 模型，可是其拟合结果很差。允许 $E \times P$ 对 M 的交互效应但仍假定 G 和 M 之间条件独立的模型拟合结果要好得多(G^2 下降了 13.0, df = 1)。进一步包括 G 的主效应的模型拟合的更好一点。后两个模型都比图 6.1 所预测的模型更复杂，因为 E 对 M 的效应随着 P 的变化而变化。然而，关于因果关系的初步思考给出的模型与最后拟合很好的模型很接近。关于第三阶段的模型效应的估计和解释，我们留给读者来完成。

表 6.4 关于表 6.3 的各模型的拟合优度[a]

阶段	结果变量	可能的解释变量	实际的解释变量	G^2	df
1	P	G	无	75.3	1
			(G)	0.0	0
2	E	G,P	无	48.9	3
			(P)	2.9	2
			$(G+P)$	0.0	1
3	M	G,P,E	$(E+P)$	18.2	5
			(E^*P)	5.2	4
			(E^*P+G)	0.7	3

a P,婚前性行为;E,婚外性行为;M,婚姻状况;G,性别。

6.1.6 数据挖掘中构建模型的新策略

随着运算能力的飞速发展,巨型的数据变得越来越平常。一个从事信用卡营销的金融机构,可能有关于上百万个他们的广告对象信用卡申请情况的观测数据。对于他们的客户,他们拥有客户是否按时付账单以及关于信用卡申请的诸多变量的月度数据。对超大规模的数据进行分析被称为*数据挖掘*(*data mining*)。

对庞大数据构建模型是很有挑战性的。目前有大量研究使用有别于传统的统计方法,包括忽视抽样误差或建模等概念的自动算法。在这种情况下,显著性检验通常是没有意义的,因为如果 n 足够大的话,任何变量几乎都会存在显著效应。即使模型具有复杂的结构,模型构建策略仍重视其对于预测的价值。不过,在将预测项加入模型的过程中,收益递减的点仍然会存在。在这一点后,新加入的变量往往与模型中已有变量的线性组合高度相关,从而不能提高模型的预测力。当 n 很大时,统计推断不如关于模型预测力的总体指标有意义。这将是下一节所要讨论的内容。

6.2 Logistic 回归诊断

在第 5.2.3 节中,我们介绍了一般意义上的检查模型拟合优度的统计量。在初步选定了一个模型后,我们转换到一种微观分析模式以求获得进一步的了解。例如,在列联表中,通过对比每个单元格的观察计数和拟合计数,由此反映的拟合不足情况可能会对进一步改进模型有所启发。对于连续的预测变量,图示法也很有价值。这些诊断分析可以告诉我们模型拟合不足的原因,如某一解释变量的非线性效应。

6.2.1 皮尔逊残差、偏离度残差和标准化残差

对于分类预测变量,计算残差来比较计数的观察值和拟合值很有帮助。令 y_i 表示在解释变量取 i 值时的 n_i 次试验对应的二项分布变量,$i=1,...,N$。令 $\hat{\pi}_i$ 表示 $P(Y=1)$ 的模型估计,那么,$n_i\hat{\pi}_i$ 就是成功数的拟合值。在一个具有二项分布随机部分的广义线性模型中,该拟合值对应的皮尔逊残差(式 4.36)为

$$e_i=\frac{y_i-n_i\hat{\pi}_i}{[\widehat{\text{var}}(Y_i)]^{1/2}}=\frac{y_i-n_i\hat{\pi}_i}{\sqrt{[n_i\hat{\pi}_i(1-\hat{\pi}_i)]}}。\tag{6.1}$$

该公式将原始残差 $(y_i-\hat{\mu}_i)$ 除以 y_i 的二项标准差的估计值。检验模型拟合优度的皮尔逊统计量满足

$$X^2 = \sum_{i=1}^{N} e_i^2,$$

即每个皮尔逊残差的平方项都是 X^2 的一部分。

将式 6.1 分子中的 $\hat{\pi}_i$ 替代为 π_i，e_i 等于二项随机变量与其期望值之差再除以其标准差的估计值。在大样本情况下，当模型成立时，e_i 近似服从 $N(0,1)$ 分布。然而，由于 π_i 是根据 $\hat{\pi}_i$ 估计的，且 $\{\hat{\pi}_i\}$ 取决于 $\{y_i\}$，因此，$\{y_i - n_i\hat{\pi}_i\}$ 往往小于 $\{y_i - n_i\pi_i\}$，$\{e_i\}$ 比 $N(0,1)$ 具有更小的方差。如果 X^2 具有 df = v，$X^2 = \sum_i e_i^2$ 渐近等同于 v 个（不是 N 个）独立的标准正态分布随机变量的平方和，那么，当模型成立时，$E(\sum_i e_i^2)/N \approx v/N < 1$。

在模型成立的情况下，标准化的皮尔逊残差的绝对值稍大，并且近似服从 $N(0,1)$。在第 4.5.5 节中，我们介绍了一种利用估计的帽子矩阵（hat matrix）中的杠杆力（leverage）进行的调整。对于杠杆力为 $\hat{h}_i$ 的第 i 个观测值，标准化的残差为

$$r_i = \frac{e_i}{\sqrt{1 - \hat{h}_i}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{[n_i\hat{\pi}_i(1 - \hat{\pi}_i)(1 - \hat{h}_i)]}}。$$

当其绝对值大于 2 或 3 时，往往表明存在拟合不足。

还有一种残差利用拟合统计量 G^2 的项，被称为*偏离度残差*（*deviance residuals*），如式 4.35 中介绍的。第 i 个观测值对应的偏离度残差可表示为

$$\sqrt{d_i} \times 符号(y_i - n_i\hat{\pi}_i), \tag{6.2}$$

其中

$$d_i = 2\left(y_i \log \frac{y_i}{n_i\hat{\pi}_i} + (n_i - y_i)\log \frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)。$$

这个残差也往往比 $N(0,1)$ 的方差小，并可以对其进行标准化。

将残差与解释变量或者线性预测值一起绘图可以反映某些拟合不足的情况。不过，当拟合值很小时，就像 X^2 和 G^2 失去意义一样，残差也变得没有意义。在解释变量为连续变量的情况下，其每一个取值处常常对应于 $n_i = 1$。这时，y_i 只能等于 0 或 1，并且 e_i 仅有两个可能的取值。因而，当结果变量的某一取值是小概率事件时，大家要特别小心，这时单个残差往往不能提供有意义的信息。同样，关于残差的绘图也只是包括两条平行线上的点而价值有限。偏离度指标也变得完全没有意义（习题 5.37）。当数据可以合并划分成一系列具有相同预测变量取值的观测值时，计算分组数据的残差要好于计算个体对象的残差。

6.2.2 例子：心脏病数据

一个住在马萨诸塞州弗雷明汉镇（Framingham，Massachusetts）的年龄为 40 ~ 59 岁的男性居民样本，按照几个变量（包括血压）进行了划分（见表 6.5）。结果变量为他们在六年的跟踪期内是否罹患冠心病。

令 π_i 表示血压水平为第 i 组的研究对象患心脏病的概率。表 6.5 给出了两个 logistic 回归模型的拟合结果及相应的标准化皮尔逊残差。第一个模型，

$$\text{Logit}(\pi_i) = \alpha,$$

该模型表示结果变量独立于血压水平，相应的残差非常大。这并不奇怪，因为这个模型拟合得很差（$G^2 = 30.0$，$X^2 = 33.4$，df = 7）。

表 6.5 对血压水平和心脏病数据拟合 Logit 模型的标准化皮尔逊残差

血压水平	样本量	发生心脏病的观察计数	拟合值		残差	
			独立性模型	线性 Logit 模型	独立性模型	线性 Logit 模型
<117	156	3	10.8	5.2	-2.62	-1.11
117-126	252	17	17.4	10.6	-0.12	2.37
127-136	284	12	19.7	15.1	-2.02	-0.95
137-146	271	16	18.8	18.1	-0.74	-0.57
147-156	139	12	9.6	11.6	0.84	0.13
157-166	85	8	5.9	8.9	0.93	-0.33
167-186	99	16	6.9	14.2	3.76	0.65
>186	43	8	3.0	8.4	3.07	-0.18

来源：数据取自：Cornfield(1962)。

对残差进行绘图，显示出上升的趋势。这表明可以考虑线性 Logit 模型，如

$$\text{Logit}(\pi_i) = \alpha + \beta x_i,$$

其中 $\{x_i\}$ 表示血压水平的赋值。这里，我们使用的赋值为（111.5,121.5,131.5,141.5,151.5,161.5,176.5,191.5）。中间的赋值等于每个血压区间的中值。该模型的残差不再具有明显趋势，且只是第二组显示出一定程度的拟合不足。

表 6.6 给出了由 SAS 计算的线性 Logit 模型的残差。皮尔逊残差(Reschi)、偏离度残差(Resdev)以及标准化皮尔逊残差(StReschi)的结果相似，每一种残差都是在第二组的值大一些。不过，某个残差值相对偏大并不奇怪。在有许多项残差的情况下，即便完全出于偶然的因素，也可能出现一些较大的值。这里，总体拟合统计量（$G^2 = 5.9$，$X^2 = 6.3$，df = 6）并没有显示任何问题。在进行残差分析时，我们应当小心不要误将模型中的随机变动当成一种趋势。

表 6.6 SAS 输出的关于表 6.5 心脏病数据的残差[a]

Observation Statistics						
Observ	disease	n	blood	Reschi	Resdev	StReschi
1	3	156	111.5	-0.979 4	-1.061 7	-1.105 8
2	17	252	121.5	2.005 7	1.850 1	2.374 6
3	12	284	131.5	-0.813 3	-0.842 0	-0.945 3
4	16	271	141.5	-0.506 7	-0.516 2	-0.572 7
5	12	139	151.5	0.117 6	0.117 0	0.126 1
6	8	85	161.5	-0.304 2	-0.308 8	-0.326 1
7	16	99	176.5	0.513 5	0.505 0	0.652 0
8	8	43	191.5	-0.139 5	-0.140 2	-0.177 3

a Reschi，皮尔逊残差；StReschi，调整后的残差。

检查拟合问题的另一种有用的图示法是比较相应比例的观察值和拟合值，将二者进行绘图或者将它们与解释变量一起绘图。图 6.2 显示了上述线性 Logit 模型中，按照血压水平划分的患心脏病的观察比例和估计概率。由图中结果来看，模型拟合得很不错。

分析残差有助于理解为什么模型存在拟合问题，或者总体拟合不错的模型在什么地方存在拟合不足。下面的例子展示了第二种情况。

6.2.3 例子：研究生录取数据

表 6.7 给出的是 1997—1998 学年佛罗里达大学人文艺术与科学学院 23 个系的研究

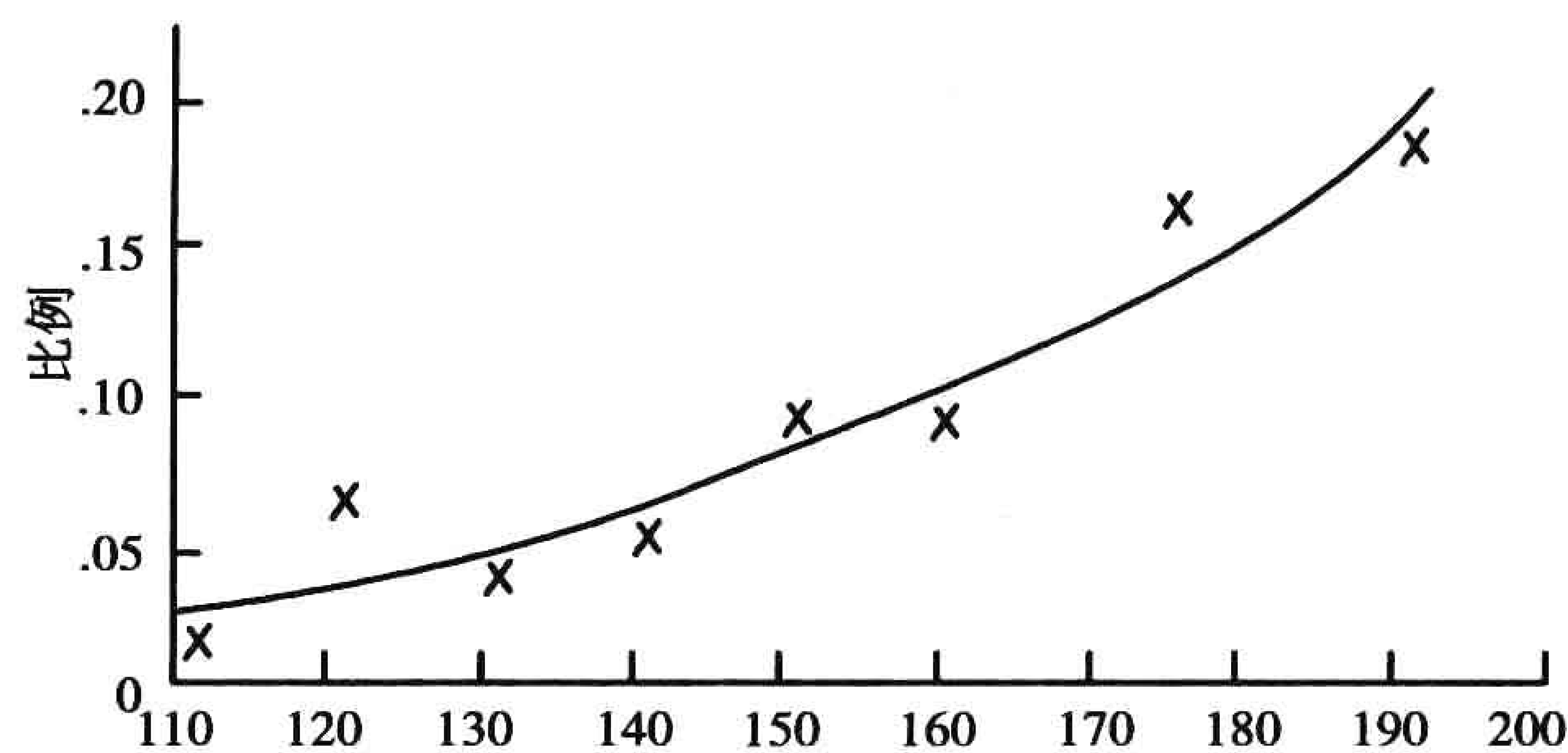


图 6.2 犯心脏病比例的观察值和线性 Logit 模型的预测值

生申请情况。按照申请人的性别(G)、是否被录取(A),以及所申请的系别(D),对数据进行了交叉分组。我们考虑以 A 为结果变量的 Logit 模型。令 y_{ik} 表示在系别 k 中性别 i 的申请人被录取的数量,π_{ik} 表示相应的录取概率。在这里,将 $\{Y_{ik}\}$ 视为服从独立的 $\text{bin}(n_{ik},\pi_{ik})$ 分布。在其他条件都相同的情况下,我们期望是否被录取的决定独立于申请人的性别。然而,给定系别,不包括性别效应的模型

$$\text{Logit}(\pi_{ik}) = \alpha + \beta_k^D,$$

拟合得非常差($G^2 = 44.7$, $X^2 = 40.9$, df = 23)。

表 6.7 关于性别、系别与研究生录取情况的数据以及不包括性别效应的模型残差

	女性		男性		标准化残差		女性		男性		标准化残差
系别	是	否	是	否	女性被录取数量	系别	是	否	是	否	女性被录取数量
anth	32	81	21	41	-0.76	ling	21	10	7	8	1.37
astr	6	0	3	8	2.87	math	25	18	31	37	1.29
chem	12	43	34	110	-0.27	phil	3	0	9	6	1.34
clas	3	1	4	0	-1.07	phys	10	11	25	53	1.32
comm	52	149	5	10	-0.63	poli	25	34	39	49	-0.23
comp	8	7	6	12	1.16	psyc	2	123	4	41	-2.27
engl	35	100	30	112	0.94	reli	3	3	0	2	1.26
geog	9	1	11	11	2.17	roma	29	13	6	3	0.14
geol	6	3	15	6	-0.26	soci	16	33	7	17	0.30
germ	17	0	4	1	1.89	stat	23	9	36	14	-0.01
hist	9	9	21	19	-0.18	zool	4	62	10	54	-1.76
lati	26	7	25	16	1.65						

来源:数据由 James Booth 友情提供。

表 6.7 同时给出了这个模型中被录取的女性数量的标准化皮尔逊残差。例如,天文学系录取了 6 名女生,比模型所预测的数量高了 2.87 个标准差。由于模型本身关于边际分布的限定,每个系仅具有一个非冗余的标准化残差。模型的拟合值为 $\hat{\pi}_{ik} = (y_{1k} + y_{2k})/n_{+k}$,对应于各分表的独立性拟合($\hat{\pi}_{1k} = \hat{\pi}_{2k}$)。这样,$y_{1k} - n_{1k}\hat{\pi}_{1k} = y_{1k} - n_{1k}(y_{1k} + y_{2k})/n_{+k} = (n_{2k}/n_{+k})y_{1k} - (n_{1k}/n_{+k})y_{2k} = -(y_{2k} - n_{2k}\hat{\pi}_{2k})$ 。因而,$(y_{1k} - n_{1k}\hat{\pi}_{1k})$ 和 $(y_{2k} - n_{2k}\hat{\pi}_{2k})$ 的标准误是相同的。男性和女性的标准化残差的绝对值相等,但符号相反。天文学系录取了 3 名男生,相应的标准化残差为 -2.87;录取的数量比预测数低了 2.87 个标准差。与普通的皮尔逊残差相比,这是标准化残差的另一个优点。在每个分表中,独立性模型的自由度为 df = 1。该标准化残差是关于数据偏离独立性模型的仅有信息,但是男性和女性所对应的普通皮尔逊残差不一定相等。

对应于较大的标准化皮尔逊残差的系揭示了模型拟合不足的原因。在天文学系和地理系,录取的女生数明显更多,而心理系则录取的女生数较少。如果不包括这三个系,模型会拟合得相对较好($G^2 = 24.4$, $X^2 = 22.8$,df = 20)。

利用完整的数据,加入性别效应并没能改进模型的拟合结果($G^2 = 42.4$, $X^2 = 39.0$, df = 22),因为上文提到的几个系与其他系相比,虽然性别的效应较大,但方向却相反。这个模型关于 *GA* 的条件发生比之比的最大似然估计值为 1.19,即在给定系别的情况下,女生被录取的发生比超过男生 19% 。相反,在不考虑系别的边际表中, *GA* 的样本发生比之比为 0.94,女生被录取的总体发生比却低 6% 。这表明,存在辛普森悖论(Simpson's paradox)的情况(第 2.3.2 节),也即条件关联与边际关联的方向相反。

6.2.4 Logistic 回归的影响力诊断

回归诊断的其他工具也可以用来评价模型的拟合情况。这些包括,将排序的残差与正态分布百分比进行绘图(Haberman 1973a)以及描述单个观测值对参数估计和拟合统计量的影响力(influence)分析。无论何种情况,当残差表明模型对某个观测值拟合得很差时,删除该观测值并利用剩下的数据重新拟合模型可以提供进一步的信息。这相当于在模型中加入一个专门针对该观测值的参数,强制对其进行完全的拟合。

与普通回归的情况一样,某个观测值可能在决定参数估计时具有较大的影响力。一个观测值的杠杆力(leverage)越大,它潜在的影响力(influence)也就越大。去除掉一个在 y 上取奇异值(outlier)并且具有很大杠杆力的观测值,模型的拟合结果可能会发生很大变化。不过,在普通回归中,一个观测值所可能具有的影响力要比在关于二分变量的logistic 回归中高得多,因为前者中 y_i 偏离其期望值的距离不存在边界。同时,正如我们在第 4.5.5节所介绍的,广义线性模型估计的帽子矩阵

$$\widehat{\mathbf{Hat}} = \hat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}^{1/2}$$

取决于拟合结果以及模型矩阵 $\mathbf{X}$。我们在第 5.5.2 节表明,在 logistic 回归中,权数矩阵 $\hat{\mathbf{W}}$ 是一个对角矩阵,对于预测变量取值为 i 时的 n_i 个观测值,它在对角线上的元素为 $\hat{w}_i = n_i\hat{\pi}_i(1 - \hat{\pi}_i)$ 。预测变量取极值的点并不一定就具有很大的杠杆力。事实上,如果 $\hat{\pi}_i$ 接近 0 或 1,杠杆力会很小。

描述从数据中删除某个观测值对参数估计和拟合统计量影响的几种测量指标在数学上与观测值的杠杆力有关(Pregibon,1981;Williams,1987)。在 logistic 回归中,观测值可能是一个单一二分结果变量的值,也可能是一组预测变量取值相同的对象的二项分布结果变量的值。每个观测值的影响力测量指标包括:

1. 对于每个模型参数,当该观测值被删除后参数估计的变动。这个变动除以它的标准误,被称为 *Dfbeta*。
2. 由于删除该观测值所引起的参数的联合置信区间的变动。这个置信区间的变动诊断工具由 c 来表示。
3. 当删除该观测值后拟合优度统计量 X^2 或 G^2 的变动。

对于以上每种测量指标,值越大就表示影响力也越大。我们以表 6.5 中用血压来预测心脏病的线性 Logit 模型为例,对此加以演示。表6.8 包括了对以下测量指标的简单近似(参见:Pregibon(1981)):关于血压系数的 *Dfbeta* 指标,置信区间诊断指标 c, G^2 的变动,以及 X^2 的变动(这是标准化皮尔逊残差的平方, r_i^2)。以上所有指标均显示,删除第二个观测值具有最大的影响。这并不奇怪,因为只有该观测值的残差相对较大。相应

地,表6.8还给出了在删除相应观测值后拟合的独立性模型中 X^2 和 G^2 的变动。在较低和较高的血压水平上,有些变动非常大。然而,这些都与删除的是某一血压水平的整个二项分布样本而不是单个对象的二分观测值有关。即便对于独立性模型,删除单个对象的影响仍然有限。

当存在连续变量或多个预测变量时,将这些诊断指标与如估计概率等进行绘图可以提供许多信息。有关诊断绘图的例子,参见:Cook and Weisberg(1999,Chap. 22),Fowlkes(1987),Landwehr et al.(1984)。

表6.8 对心脏病数据拟合的 Logistic 回归模型的诊断指标

血压	*Dfbeta*	c	皮尔逊 X^2 之差	似然比 G^2 之差	皮尔逊 X^2 之差[a]	似然比 G^2 之差[a]
111.5	0.49	0.34	1.22	1.39	6.86	9.13
121.5	-1.14	2.26	5.64	5.04	0.02	0.02
131.5	0.33	0.31	0.89	0.94	4.08	4.56
141.5	0.08	0.09	0.33	0.34	0.55	0.57
151.5	0.01	0.00	0.02	0.02	0.70	0.66
161.5	-0.07	0.02	0.11	0.11	0.87	0.80
176.5	0.40	0.26	0.42	0.42	14.17	10.83
191.5	-0.12	0.02	0.03	0.03	9.41	6.73

a 独立性模型;其他值是指以血压作为预测变量的模型。

来源:数据取自:Cornfield(1962)。

6.2.5 总体预测力:R 和 R^2 指标

在普通回归中,R^2 描述与边际方差相比结果变量的条件方差的消减比例。R^2 和多元相关系数 R 表示解释变量对结果变量的预测力,其中 $R=1$ 表示完全预测。尽管有许多人尝试界定适用于分类结果变量的类似指标,但是没有一种指标能像 R 和 R^2 那样得到广泛的应用。在这一节中,我们介绍几种相应的指标。

对于所有的广义线性模型,结果变量的观测值 $\{y_i\}$ 和模型的拟合值 $\{\hat{\mu}_i\}$ 之间的相关系数 $r(y,\hat{\mu})$ 可以用来描述预测力。在最小二乘法回归的情况下,这就是 Y 和预测变量之间的多元相关系数。相关系数相对于其平方的优点在于,它使用的是初始刻度并且它与效应的大小近似成比例。当某一预测变量具有较小的效应时,将其斜率的取值翻倍大致对应着相关系数也翻倍。该指标可用来比较关于同一数据的不同模型的拟合情况。

在 logisitc 回归中,某个特定模型的 $\hat{\mu}_i$ 是关于二分观测值 i 的估计概率 $\hat{\pi}_i$。表6.2给出了利用马蹄蟹数据拟合的几个模型的 $r(y,\hat{\mu})$。只包括壳宽的模型中 $r=0.402$,在模型中加入颜色后 r 上升到0.452。仅仅区分颜色是否为黑色的较简单模型基本上拟合得一样好,$r=0.447$。包括颜色、蟹刺状况、壳宽以及它们之间所有的二维和三维交互项的复杂模型具有 $r=0.526$。这一系数看上去好像明显更高,但在具有多个预测变量的情况下,利用 r 来估计真实的相关系数会变得极不准确。对于自由度相差很多的模型,比较它们的 r 值可能会得出误导性的结论。在经过一种旨在降低偏误的刀切法(jackknife)调整之后,这个过于复杂的模型与较简单模型所对应的 r 几乎相同(Zheng and Agresti ,2000)。因而,使用较为简单的模型的所得远远大于所失。

度量二分结果变量 $\{y_i\}$ 及其拟合值 $\{\hat{\pi}_i\}$ 之间的关联的另一种方法是利用平方差的消减比例

$$1-\frac{\sum_i (y_i-\hat{\pi}_i)^2}{\sum_i (y_i-\bar{y})^2},$$

即使用 $\hat{\pi}_i$ 而不是 $\bar{y}=\sum y_i/n$ 来预测 y_i 时所少犯的错误(Efron,1978)。Amemiya(1981)建议使用预测方差的倒数对标准差的平方进行加权。然而,与正态的广义线性模型不同,随着模型中的参数增加,logistic 回归中这些指标以及 $r(y,\hat{\mu})$ 有可能会下降。与其他的相关系数形式的指标一样,它们强烈依赖于观测值中解释变量的取值范围。

还有其他的指标直接使用似然函数。记某一给定模型的最大对数似然值为 L_M,饱和模型(saturated model)和只包括截距项的空模型(null model)的最大对数似然值分别为 L_S 和 L_0。由于概率的值不能超过 1.0,对数似然值必然非正。随着模型变得更复杂,参数空间进一步扩张,最大的对数似然值会上升。因此,$L_0 \leqslant L_M \leqslant L_S \leqslant 0$。指标

$$\frac{L_M-L_0}{L_S-L_0} \tag{6.3}$$

的取值在 0 到 1 之间。当模型的拟合相对于空模型没有任何改进时,该指标等于 0;当模型的拟合与饱和模型一样好时,该指标等于 1。此指标的一个缺点在于,对数似然值并不是一个易于解释的测度。除用于比较不同的模型外,很难对指标的取值本身进行解释。

对于 n 个独立的伯努利观测值,最大化的对数似然值等于

$$\log \prod_{i=1}^{n} [\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}] = \sum_{i=1}^{n} [y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)]。$$

在空模型中存在 $\hat{\pi}_i = (\sum y_i)/n = \bar{y}$,所以

$$L_0 = n[\bar{y}(\log \bar{y}) + (1-\bar{y})\log(1-\bar{y})]。$$

饱和模型中每个对象都对应一个参数,这意味着,对所有的 i 都存在 $\hat{\pi}_i = y_i$。因而,$L_S = 0$,式 6.3 可简化为

$$D = \frac{L_0 - L_M}{L_0}。$$

这个度量指标是由 McFadden(1974)提出的。

当在解释变量的每个取值处都存在多个观测值时,数据文件可以表示为 N 个二项分布计数的分组数据形式,而不是 n 个伯努利观测值。这时,饱和模型针对每个计数都存在单独的参数。它使得 N 个拟合比例等于 N 个结果为成功的样本比例。在这种情况下,L_S 不等于零,并且式 6.3 所取的值与通过个体对象所计算的值不同。对于 N 个二项分布计数,最大化的似然值与 G^2 拟合优度统计量有关,存在 $G^2(M) = -2(L_M - L_S)$,因而式 6.3 对应于

$$D^* = \frac{G^2(0) - G^2(M)}{G^2(0)}。$$

关于该指标以及相应的偏关联指标的讨论,参见:Goodman(1971a)、Theil(1970)。

在分组数据的情况下,即便模型在个体对象层面上的预测力很弱,D^* 的值也可能很大。例如,即使对于整个样本而言拟合概率都接近 0.5,该模型也可以比空模型拟合得好很多。尤其是,当模型出现完全拟合时,就有 $D^* = 1$,不论该模型到底能在多大程度上预测个体对象的 Y 值。另外,假定某个给定模型对于总体成立,但空模型不成立,在组数 N 不变的情况下,随着样本规模 n 的增加,$G^2(M)$ 为一个随机的卡方变量,但 $G^2(0)$ 会无限增大。因而,当 $n \to \infty$ 时,$D^* \to 1$,并且它的大小会随着 n 的变动而变动。该指标将模型的拟合优度与预测力相混淆。在普通回归分析中,当使用在 x 的不同取值水平上的

N 个 Y 的均值(而不是个体对象)来计算 R^2 时,也会发生这种问题。因而,还是对未分组的二分数据使用 D 更合理一些。

6.2.6 总体预测力:分类表和 ROC 曲线

分类表(*classification table*)将二分结果变量与关于 $y = 0$ 或 1 的预测结果进行了交叉分组。在预测中,对于某个切点 π_0 ,当 $\hat{\pi}_i > \pi_0$ 时 $\hat{y} = 1$,当 $\hat{\pi}_i \leqslant \pi_0$ 时 $\hat{y} = 0$ 。大多数分类表使用 $\pi_0 = 0.5$,并通过以下指标来反映预测力:

$$灵敏度 = P(\hat{y} = 1 \mid y = 1) \quad 以及 \quad 准确度 = P(\hat{y} = 0 \mid y = 0)$$

(回顾第 2.1.2 节)。分类表方法的局限在于,它将连续的预测值 $\hat{\pi}$ 合并为两个值,π_0 的选取具有随意性,并且它对 $y = 1$ 和 $y = 0$ 的相对发生频数非常敏感。

接收机工作特性(*receiver operating characteristic*,*ROC*)曲线是对于可能的切点 π_0 ,将灵敏度(sensitivity)作为(1 - 准确度(specificity))的函数的绘图。它通常是一条连接点(0,0)和(1,1)的凸曲线。曲线下面的面积越大,模型的预测就越好。ROC 曲线比分类表提供了更多的信息,因为它给出了所有可能的 π_0 所对应的预测力。图 6.3 显示了有关马蹄蟹数据的模型的 ROC 曲线(由 SAS 的 PROC LOGISTIC 程序绘制),以壳宽和颜色作为预测变量。

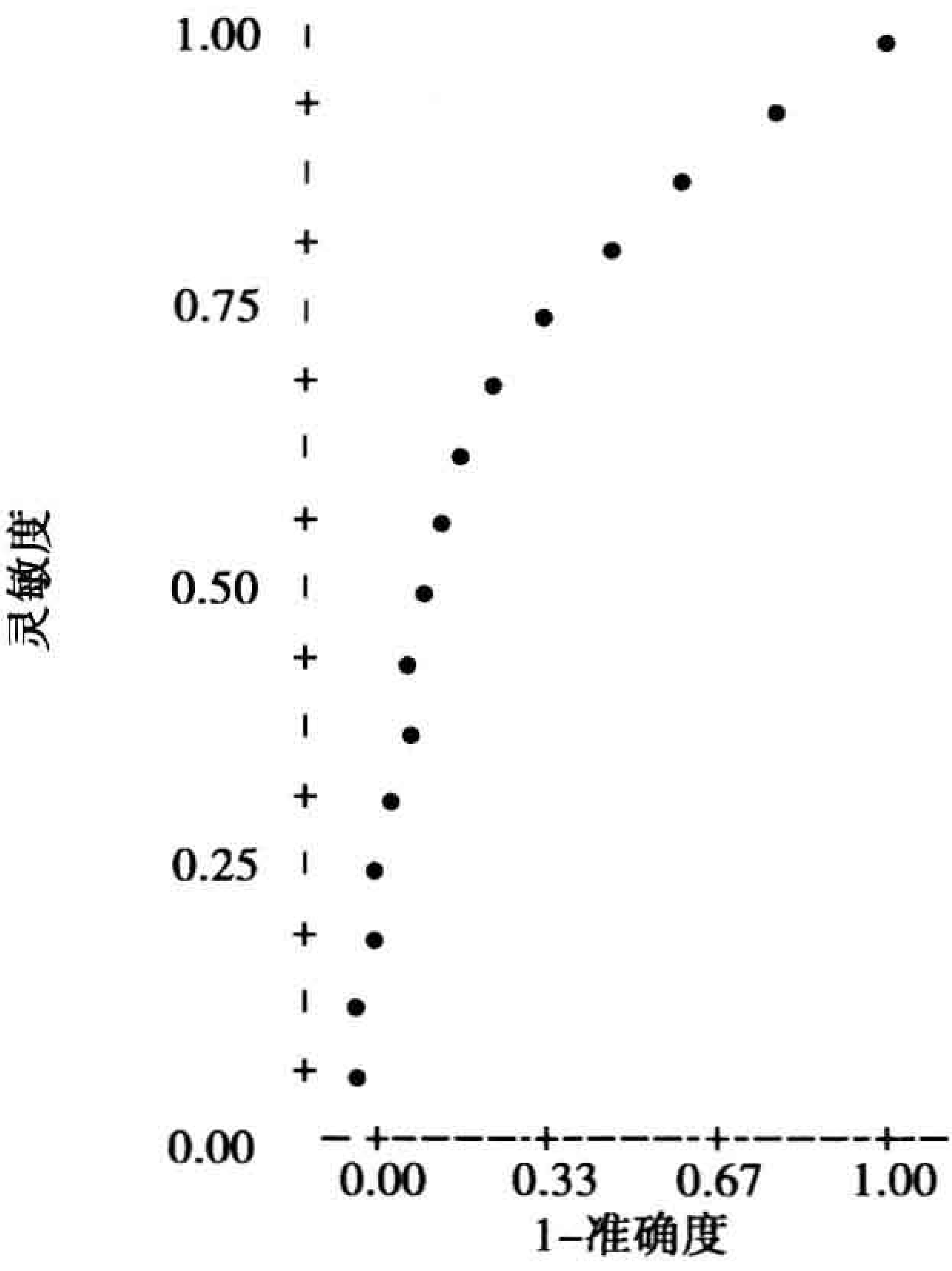


图 6.3 关于马蹄蟹数据的 logistic 回归模型的 ROC 曲线

ROC 曲线下面的面积等于另一个预测力测量指标,*即相协指数*(*concordance index*)的取值。考虑所有观测值的配对 (i,j) ,比如 $y_i = 1$ 和 $y_j = 0$,相协指数 c 估计预测情况与结果相一致的概率。y 的取值较大的观测值,其对应的 $\hat{\pi}$ 也较大(Harrell et al. ,1982)。$c = 0.5$ 表示预测情况并不比随机猜测的结果更好。它对应着一个仅包括截距项的模型,这时的 ROC 曲线是一条连接点(0,0)和(1,1)的直线。对于马蹄蟹数据,当颜色作为唯一的预测变量时 $c = 0.639$,壳宽作为唯一预测变量时 $c = 0.742$,将壳宽和颜色都作为预测变量时 $c = 0.771$,当使用壳宽以及表示颜色是否为黑色的虚拟变量时 $c = 0.772$ 。

ROC 曲线是一种常见的模型评估诊断工具。有时,这种检验的结果变量具有 $J > 2$ 种有序的类别而不仅仅是(正,负)两类。这时,对结果为正的可能切点,ROC 曲线会存在多种定义,它是对 J 个类别在各种可能的(正,负)划分下的灵敏度和(1 - 准确度)进行的绘图(参见:Toledano and Gatsonis(1996))。

6.3 $2 \times 2 \times K$ 表格中条件关联的统计推断

在第6.2.3节中,我们使用条件独立性模型分析了研究生录取数据。该模型在生物医药研究中非常重要,常用以考察在控制了可能的混淆变量(confounding variable)后,干预变量和疾病结果的关联是否仍然存在。在本节中,我们讨论对 $2 \times 2 \times K$ 列联表的 Logit 模型分析的条件独立性检验。此外,我们还介绍一个看似非模型的,但与 Logit 模型有关的检验(Mantel and Haenszel,1959)。

我们利用表6.9中一项关于8个中心的临床试验(clinical trial)结果对此加以展示。该研究比较两种乳膏制剂——一种药品和一种控制品——治疗某种感染的效果。该表反映了一种常见的药学应用,即通过取自不同层(strata)的观测值来比较两种干预方式对一个二分结果变量的影响。这里的层通常是指医学中心或诊所,也可以由年龄组或疾病的严重程度甚至几个控制变量的取值组合来构成;此外,在元分析(meta analysis)中,层也可以表示性质相同的不同研究。

表6.9 8个中心临床试验的干预方式与结果

中心	干预方式	结果		发生比之比	μ_{11k}	$\mathrm{var}(n_{11k})$
		有效	无效			
1	药物	11	25	1.19	10.36	3.79
	控制	10	27			
2	药物	16	4	1.82	14.62	2.47
	控制	22	10			
3	药物	14	5	4.80	10.50	2.41
	控制	7	12			
4	药物	2	14	2.29	1.45	0.70
	控制	1	16			
5	药物	6	11	∞	3.52	1.20
	控制	0	12			
6	药物	1	10	∞	0.52	0.25
	控制	0	10			
7	药物	1	4	2.0	0.71	0.42
	控制	1	8			
8	药物	4	2	0.33	4.62	0.62
	控制	6	1			

来源:Beitler and Landis(1985)。

6.3.1 通过 Logit 模型检验条件独立性

对于二分结果变量 Y,我们研究在控制了分类协变量(covariate) Z 后二分预测变量 X 的效应。令 $\pi_{ik} = P(Y = 1 \mid X = i, Z = k)$,考虑模型

$$\mathrm{Logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad i = 1,2, \quad k = 1,\cdots,K, \tag{6.4}$$

其中 $x_1 = 1$, $x_2 = 0$。这个模型假定在 Z 的每个类别中,XY 的条件发生比之比都相等,即等于 $\exp(\beta)$。XY 条件独立性的零假设为 $H_0: \beta = 0$。相应的沃尔德统计量为 $(\hat{\beta}/SE)^2$。似然比统计量等于简约模型(reduced model)

$$\mathrm{Logit}(\pi_{ik}) = \alpha + \beta_k^Z \tag{6.5}$$

与全模型(full model)的 G^2 统计量之差。当 X 的效应在 Z 的每个类别中都相似时,这些检验是适当的。检验的自由度为 df = 1。

同样地,由于简约模型(式6.5)等价于 X 和 Y 条件独立的情况,可以利用该模型的拟合优度检验来检验条件独立性。当 X 是二分变量时,相应检验具有自由度 df = K。这对应于将式6.5 模型与饱和模型进行比较,后者允许 $\beta \neq 0$ 并包括交互效应 XZ 的参数。当不存在交互效应或者交互效应存在但不具备实质意义时,根据后面将要介绍的结果(第6.4.2 节),这种方法的效能较差,尤其是在 K 很大的情况下。然而,当 XY 关联的方向随着 Z 的类别变化而改变时,这种方法却具有更大的效能。

6.3.2 条件独立性的 Cochran-Mantel-Haenszel 检验

针对 $H_0: 2\times2\times K$ 表格中的条件独立性,Mantel 和 Haenszel(1959)提出了一种不基于模型的检验。在有关疾病的回顾性研究中,他们将结果变量(列)的边际总计视为给定的。因此,在单元格计数为 $\{n_{ijk}\}$ 的第 k 个分表中,他们的分析以预测变量总计 (n_{1+k}, n_{2+k}) 和结果变量总计 (n_{+1k}, n_{+2k}) 为条件。这时,在常用的抽样方案中,每个分表的第一个单元格计数 n_{11k} 服从超几何分布(式3.16)。给定边际的总计,这个计数决定了 $\{n_{12k}, n_{21k}, n_{22k}\}$ 。

在 H_0 下, n_{11k} 的超几何分布均值和方差分别为:

$$\mu_{11k} = E(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n_{++k}},$$

$$\text{var}(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}。$$

不同分表中的单元格计数是相互独立的。通过比较 $\sum_k n_{11k}$ 和相应的零期望值,检验统计量合并了 K 个分表的信息。它等于

$$\text{CMH} = \frac{\left[\sum_k (n_{11k} - \mu_{11k})\right]^2}{\sum_k \text{var}(n_{11k})}。\tag{6.6}$$

该统计量服从自由度 df = 1 的大样本卡方零分布。

当第 k 个分表的发生比之比 $\theta_{XY(k)} > 1$ 时,我们期望 $(n_{11k} - \mu_{11k}) > 0$ 。当每个分表中都有 $\theta_{XY(k)} > 1$ 或 $\theta_{XY(k)} < 1$ 时, $\sum_k (n_{11k} - \mu_{11k})$ 的绝对值一般会比较大。在每个分表中的 XY 关联都相似时,这个检验效果最好。从这个意义上说,它与在 Logit 模型(式6.4)中检验 $H_0: \beta = 0$类似。当各层的样本规模比较大时,这个检验通常会给出与模型检验相似的结果。事实上,它是在模型中对 $H_0: \beta = 0$进行的计分检验(第 1.3.3 节)(Day and Byar ,1979)。

Cochran(Cochran,1954)提出了一个相似的统计量。他将每个 2 × 2 表格中的行看作两个独立的二项分布样本而不是一个超几何分布样本。Cochran 的统计量是将式 6.6 中的 $\text{var}(n_{11k})$ 替代为

$$\text{var}(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^3}。$$

由于这两种方法很相似,我们将式6.6 称为 *Cochran-Mantel-Haenszel(CMH)统计量*(*Cochran-Mantel-Haenszel statistic*)。Mantel 和 Haenszel 使用超几何分布的方法更具有一般性,因为它也适用于行不是两个总体的独立二项分布样本的情况。这样的例子包括回顾性研究以及研究对象被随机分配到两个干预组的随机临床试验(randomized clinical trial)。在前一种情况下,列总计是自然给定的。后一种情况中,在零假设下无论研究对象如何进行分组,列边

际都是相同的,并且随机化分组导致每个 2×2 表格都是超几何分布样本。

Mantel 和 Haenszel(1959)提出了一个对式 6.6 的连续性修正(continuity correction)。这使得检验的 P 值能够更好地近似于精确条件检验(第 6.7.5 节),但其结果往往偏于保守。CMH 统计量可以扩展到 $I \times J \times K$ 表格的情况(第 7.5.3 节)。

6.3.3 例子:多中心临床试验

对于多中心临床试验数据,表 6.9 给出了每个分表的样本发生比之比,以及在 H_0: 条件独立性下药物干预成功数量(n_{11k})的期望值和方差。除最后一个分表外,其他分表中的样本发生比之比都显示正向关联。因此,我们可以合并这些结果,CMH = 6.38, df = 1。检验给出了拒绝 H_0 ($P = 0.012$)的明确证据。

在 Logit 模型(式 6.4)中检验 H_0: $\beta = 0$得到的结论与此相似。模型拟合结果中 $\hat{\beta} = 0.777$,标准误为 SE $= 0.307$ 。相应的沃尔德统计量等于 $(0.777/0.307)^2 = 6.42$ ($P = 0.011$),似然比统计量等于 6.67($P = 0.010$)。

6.3.4 CMH 检验与稀疏数据*

总之,对于 Logit 模型(式 6.4),CMH 是关于 H_0: $\beta = 0$的似然比或沃尔德检验所对应的计分检验。给定 K 固定不变,当 $n \rightarrow \infty$ 时,在 H_0 下这些检验渐近服从相同的卡方分布。CMH 的一个优点是,随着 $n \rightarrow \infty$,它的卡方极限在 $K \rightarrow \infty$ 的渐近情况下也适用。关于似然比和沃尔德检验的渐近理论则要求参数的数量(因而 K)是给定的,所以它们不适用于上述情况。在实际应用中,当每层只包括一个配对,即每组只有一个对象时,就会出现这种情况。

在只包括一个配对的层中,对于每个 k 存在 $n_{1+k} = n_{2+k} = 1$ 。因此, $n = 2K$,当 $n \rightarrow \infty$ 时, $K \rightarrow \infty$ 。表 6.10 展示了这种情况下的数据结构。当第 k 层中两个对象的结果变量取值相同时(如表 6.10 的第一种情况), $n_{+1k} = 0$ 或 $n_{+2k} = 0$ 。给定边际计数,内部的计数就全部确定了,并有 $\mu_{11k} = n_{11k}$ 以及 $\mathrm{var}(n_{11k}) = 0$ 。当对象间的结果变量取值不同时(如第二种情况), $n_{+1k} = n_{+2k} = 1$,这时, $\mu_{11k} = 0.5$ 且 $\mathrm{var}(n_{11k}) = 0.25$ 。因此,只有当两个对象的结果变量取值不同时,这个配对才会影响 CMH 统计量。令 K^* 表示 K 个分表中满足这一条件的数量。尽管每个 n_{11k} 仅可以取两个值,中心极限定理表明,在 K^* 很大的情况下, $\sum_k n_{11k}$ 近似于正态分布。因而,CMH 近似于卡方分布。

通常来说,当 K 随着 n 增大而增大时,每个层中的观测值数量都较少。当然,每个层有可能包括多于两个的观测值,比如在个案-控制研究中对一个个案匹配多个控制案例的情况。观测值相对较少的列联表被归结为*稀疏*(*sparse*)表格。当 $n \rightarrow \infty$ 时 $K \rightarrow \infty$ 的特殊情形被称为*稀疏数据渐近特性*(*sparse-data asymptotics*)。由于参数的数量不固定并随着样本规模的变动而变动,普通的最大似然估计无法实现。特别地,只有当多数层的边际总计超过 5 到 10、K 是给定的且相对于 n 较小时,似然比和沃尔德统计量才近似服从卡方分布。

表 6.10　包含一个配对的层

配对的元素	结果		结果	
	成功	失败	成功	失败
第一个	1	0	1	0
第二个	1	0	0	1

6.3.5 共同发生比之比的估计

估计关联的强度比对它进行假设检验更具价值。当各分表中的关联看上去很稳定时,将 K 个样本发生比之比合并成一个条件关联的综合指标很有意义。式 6.4Logit 模型隐含着同质性关联(homogeneous association),即 $\theta_{XY(1)} = \cdots = \theta_{XY(K)} = \exp(\beta)$ 。这时,共同发生比之比(common odds ratio)的最大似然估计等于 $\exp(\hat{\beta})$ 。

其他关于共同发生比之比的估计是不基于模型的。Woolf(1955)提出了一种对 K 个样本发生比之比进行指数加权平均的方法。Mantel 和 Haenszel(1959)提出了如下指标:

$$\hat{\theta}_{\mathrm{MH}} = \frac{\sum_k (n_{11k}n_{22k}/n_{++k})}{\sum_k (n_{12k}n_{21k}/n_{++k})} = \frac{\sum_k p_{11|k}p_{22|k}n_{++k}}{\sum_k p_{12|k}p_{21|k}n_{++k}}, \tag{6.7}$$

其中 $p_{ij|k} = n_{ijk}/n_{++k}$ 。它对样本规模较大的层赋予了更大的权重。当 K 比较大并且存在数据稀疏问题时,这种方法优于最大似然估计值。这种情况下,关于对数发生比之比的最大似然估计值 $\hat{\beta}$ 的绝对值一般会偏大。例如,对于每个层只包括一个配对的情况,稀疏数据渐近特性存在 $\hat{\beta} \xrightarrow{p} 2\beta$ (这种*依概率收敛*(*convergence in probability*)的意思是,对于任意 $\varepsilon > 0$,随着 $n \to \infty$,$P(|\hat{\beta} - 2\beta| < \varepsilon) \to 1$;参见习题 10.24)。

Hauck(1979)给出了在层数固定情况下关于 $\log(\hat{\theta}_{\mathrm{MH}})$ 的渐近方差。这时,除非 $\beta = 0$,$\log(\hat{\theta}_{\mathrm{MH}})$ 的统计效力比最大似然估计值 $\hat{\beta}$ 略差(Tarone et al.,1983)。Robins 等(1986)推导出了一个既适用于 n 很大但 K 固定的标准渐近特性,又适用于 K 也很大的稀疏渐近特性的估计方差。令 $\hat{\theta}_{\mathrm{MH}} = R/S = (\sum_k R_k)/(\sum_k S_k)$,其中 $R_k = n_{11k}n_{22k}/n_{++k}$,他们的推导证明,$(\log\hat{\theta}_{\mathrm{MH}} - \log\theta)$ 与 $(R - \theta S)$ 近似成比例。同时,他们还给出 $E(R - \theta S) = 0$,并推导了 $(R - \theta S)$ 的方差,其结果为

$$\begin{aligned}\hat{\sigma}^2[\log\hat{\theta}_{\mathrm{MH}}] = {} & \frac{1}{2R^2}\sum_k n_{++k}^{-1}(n_{11k} + n_{22k})R_k + \\ & \frac{1}{2S^2}\sum_k n_{++k}^{-1}(n_{12k} + n_{21k})S_k + \\ & \frac{1}{2RS}\sum_k n_{++k}^{-1}[(n_{11k} + n_{22k})S_k + (n_{12k} + n_{21k})R_k]。\end{aligned}$$

对于表 6.9 中的八个中心临床试验数据,

$$\hat{\theta}_{\mathrm{MH}} = \frac{(11 \times 27)/73 + \cdots + (4 \times 1)/13}{(25 \times 10)/73 + \cdots + (2 \times 6)/13} = 2.13 。$$

由此可得,$\log\hat{\theta}_{\mathrm{MH}} = 0.758$,$\hat{\sigma}[\log\hat{\theta}_{\mathrm{MH}}] = 0.303$ 。关于共同发生比之比的 95% 的置信区间为 $\exp(0.758 \pm 1.96 \times 0.303)$,即 $(1.18, 3.87)$ 。式 6.4 模型也得出了相似的结果。根据沃尔德方法,关于 $\exp(\beta)$ 的 95% 的置信区间等于 $\exp(0.777 \pm 1.96 \times 0.307)$,即 $(1.19, 3.97)$ 。利用似然比方法所得出的相应区间为 $(1.20, 4.02)$ 。尽管这些结果给出了很强的存在某种效应的证据,但是关于效应大小的推断却很不精确。使用该药品后病情缓解的发生比可能只提高了 20% ,也可能提高了四倍。

如果分表中的真实发生比之比不相等但是变动不是很大,$\hat{\theta}_{\mathrm{MH}}$ 仍然是描述条件关联的一个有用指标。类似地,只要样本的关联方向基本一致,CMH 检验就是对 H_0: 条件独立性进行假设检验的非常有效的工具。在使用 CMH 检验时,并不需要假定所有的发生比之比都相等。

6.3.6 检验发生比之比的同质性

对于 $2 \times 2 \times K$ 表格,同质性关联的条件 $\theta_{XY(1)} = \cdots = \theta_{XY(K)}$ 等价于式6.4Logit 模型。对同质性关联的检验隐含着对该模型拟合优度的检验。常用的 G^2 和 X^2 检验统计量就是如此,其自由度为 $df = K - 1$。它们检验饱和模型中表示交互项(虚拟变量 x 和代表 Z 的类别的 $(K-1)$ 个虚拟变量的交互乘积项)系数的 $K-1$ 个参数都等于0。对此,Breslow 和 Day(1980,p. 142)给出了另一种大样本检验方法(注解6.5)。

在表6.9中的8中心临床试验数据中,$G^2 = 9.7$,$X^2 = 8.0$ (df=7),检验结果与发生比之比相等的假设并不矛盾。这时,用一个单一的发生比之比(例如,$\hat{\theta}_{MH} = 2.1$)来综述所有八个分表中的条件关联是适当的。事实上,即使同质性关联检验的 P 值很小,如果样本发生比之比的变动性不是很明显,诸如 $\hat{\theta}_{MH}$ 等综合性指标仍然非常有用。因而,不能把同质性检验当作使用该指标或检验条件独立性的必要前提。

6.3.7 发生比之比的异质性问题

在实际应用中,预测变量的效应在各层间往往是相似的。例如,在对一种新药品和标准药品进行比较的多中心临床试验中,如果该新药品确实效果更好,它的真实效应一般在每个层中都会是正的。

但是,严格来说,具有同质性效应的模型是不现实的。首先,我们很少会期待各层间的真实发生比之比是完全相等的,因为它会受到一些未测量的协变量的影响。Breslow(1976)利用一系列解释变量讨论了关于对数发生比之比的模型分析。第二,模型将层效应 $\{\beta_k^Z\}$ 视为固定效应(fixed effects),将这些层看作唯一关注的对象。但是,这些层往往是代表所有可能的层的一个样本。多中心临床试验包括了某些中心的数据,但是其他许多中心也可能做这类的研究。研究者希望他们的研究结论可以适用于类似的所有中心,而不仅仅是研究中所包括的那些。

另一种 Logit 模型将分表中的真实对数发生比之比视为一个服从 $N(\mu,\sigma^2)$ 分布的随机变量。通过拟合该模型,可以得出对数发生比之比的估计均值及其方差估计。这种推断适用于所有层的总体,而不仅仅是那些被选中的层。该模型通过在线性预测项中加入*随机效应*(*random effects*)来涵盖这种额外的变动性。关于包括随机效应的广义线性模型,我们将在第12章讨论,并在第12.3.4节对表6.9的数据拟合随机效应模型。

6.4 利用模型提高推断效能

我们在第3.4节指出,当列联表由定序变量组成时,利用排序信息进行检验具有更强的统计效能。在线性 Logit 模型(第5.3.4和第5.4.6节)中,以线性趋势作为备择假设的独立性检验可以达到这一目标。在此,我们讨论这样做能够提高统计效能的原因。

6.4.1 定向的备择假设

考虑由 I 个参数分别为 $\{\pi_i\}$ 的二项分布变量所形成的 $I \times 2$ 列联表,H_0:独立性可表述为

$$\text{Logit}(\pi_i) = \alpha 。$$

第3.2.1节中介绍的普通 X^2 和 G^2 统计量所对应的一般性备择假设为

$$\text{Logit}(\pi_i) = \alpha + \beta_i ,$$

这是饱和模型。在该模型中，X^2 和 G^2 检验 $H_0:\beta_1=\beta_2=\cdots=\beta_I=0$，自由度为 df =（I－1）。这种一般性备择假设将类别的划分都视为定类尺度。我们将这些检验统计量分别表示为 $G^2(I)$ 和 $X^2(I)$ 。回想一下，$G^2(I)$ 就是比较饱和模型 M_1 和独立性（ I ）模型 M_0 的似然比统计量 $G^2(M_0 \mid M_1) = -2(L_0 - L_1)$ 。

定序的检验统计量针对的是范围较窄、通常来说更为相关的备择假设。在行变量是定序变量的情况下，一个这样的例子是在线性 Logit 模型—— $\text{Logit}(\pi_i) = \alpha + \beta x_i$ 中检验 $H_0:\beta=0$。似然比统计量 $G^2(I\mid L) = G^2(I) - G^2(L)$ 用来比较线性 Logit 模型和独立性模型。当检验统计量考察单个参数时，如上述模型中的 β，它具有自由度 df = 1。在这里，检验自由度等于卡方分布的均值。给定检验统计量的取值，$X^2(I)$ 或 $G^2(I)$ 在 df = 1 时比 df = $(I-1)$时落在卡方分布右边的更远处。因此，df = 1 时对应的 P 值更小。

6.4.2 非中心卡方分布

为了比较 $G^2(I\mid L)$ 和 $G^2(I)$ 的效能，有必要比较它们的非零样本分布。当 H_0 不成立时，它们的分布近似于*非中心卡方分布*（*noncentral chi-squared*）。这个分布最早由 R. A. Fisher 在 1928 年提出，它可以表示为：如果 $Z_i \sim N(\mu_i,1)$ ，$i = 1,\cdots,v$ ，且 $Z_1,\cdots,Z_v$ 相互独立，那么 $\sum Z_i^2$ 服从 df = v 的非中心卡方分布，其*非中心参数*（*noncentral parameter*）$\lambda = \sum \mu_i^2$ 。它的均值等于 $v+\lambda$ ，方差等于 $2(v+2\lambda)$ 。普通的（中心）卡方分布具有 $\lambda = 0$ ，它对应于 H_0 为真的情况。

令 $X^2_{v,\lambda}$ 表示 df = v 、非中心参数为 λ 的非中心卡方随机变量。关于卡方分布的一个基本结论是，对于给定的 λ ，

$$P[X^2_{v,\lambda} > X^2_v(\alpha)] \text{ 随着 } v \text{ 的下降而增大。}$$

也就是说，在一个给定的 α 水平上拒绝 H_0 的效能随着检验自由度的减小而上升（如 Das Gupta and Perlman，1974）。对于给定的 v ，当 $\lambda = 0$ 时，效能等于 α ，并且它随着 λ 的上升而上升。效能与自由度之间的负相关意味着，将非中心性集中于一个具有较小自由度的统计量可以提高效能。

6.4.3 较窄的备择假设所提高的效能

假定 X 对 $\text{Logit}[P(Y=1)]$ 至少具有近似的线性效应。为了检验独立性，往往需要用到一个对该效应有很强效能的统计量。这就是通过线性 Logit 模型进行检验的目的，包括运用似然比统计量 $G^2(I|L)$、沃尔德统计量 $z = \hat{\beta}/SE$ 以及 Cochran-Armitage（计分）统计量。

那么，在什么情况下 $G^2(I\mid L)$ 会比 $G^2(I)$ 统计效能更高呢？这两个统计量满足

$$G^2(I) = G^2(I\mid L) + G^2(L) ,$$

其中 $G^2(L)$ 检验线性 Logit 模型的拟合优度。当线性 Logit 模型成立时，$G^2(L)$ 渐近服从 df = $I-2$ 的卡方分布；当 $\beta \neq 0$ 时，$G^2(I)$ 和 $G^2(I\mid L)$ 都近似服从具有相等的非中心参数的非中心卡方分布。然而，$G^2(I)$ 的自由度为 df = $I-1$ ，$G^2(I\mid L)$ 的自由度为 df = 1。因此，$G^2(I\mid L)$ 的效能更高，原因在于它使用了更少的自由度。

当线性 Logit 模型不成立时，$G^2(I)$ 的非中心参数会比 $G^2(I\mid L)$ 的大，模型拟合得越差，它们之间的差异也就越大。然而，当模型对现实情况的近似基本可以接受时，通常

$G^2(I \mid L)$ 的统计效能仍然更强。检验自由度为 1 的收益完全可以补偿它在非中心参数上的损失。真实的关系越接近于线性 Logit，$G^2(I \mid L)$ 的非中心参数就越接近于 $G^2(I)$，进而前者比后者的效能也越高。为了演示这一结论，图 6.4 显示当自由度分别等于 1 和 7 时效能作为非中心参数的一个函数。当 df = 1 的检验的非中心参数至少是 df = 7 时的一半时，df = 1 的检验具有更强的效能。因此，线性 Logit 模型有助于发现关联的一个关键组成部分。正如 Mantel(1963)在类似语境下所指出的，"检验一个线性回归并不意味着我们做了线性假定。相反，对回归的线性组成部分的检验，为发现进一步的关联提供了效能"。

这里增进的效能是以牺牲其他情况下的效能为代价的。当线性 Logit 模型对现实的描述非常糟糕时，$G^2(I)$ 检验会比 $G^2(I \mid L)$ 具有更高的效能。

有关非中心参数的结论也适用于定类变量。例如，考虑一个 $2 \times 2 \times K$ 表格的条件独立性检验。一种方法是在式 6.4 模型中检验 $\beta = 0$，相应的自由度 df = 1。另一种方法检验式 6.5 模型的拟合优度，自由度为 df = K(第 6.3.1 节)。当式 6.4 模型成立时，两个检验具有相同的非中心参数。因而，关于 $\beta = 0$ 的检验效能更高，因为它使用的自由度更少。

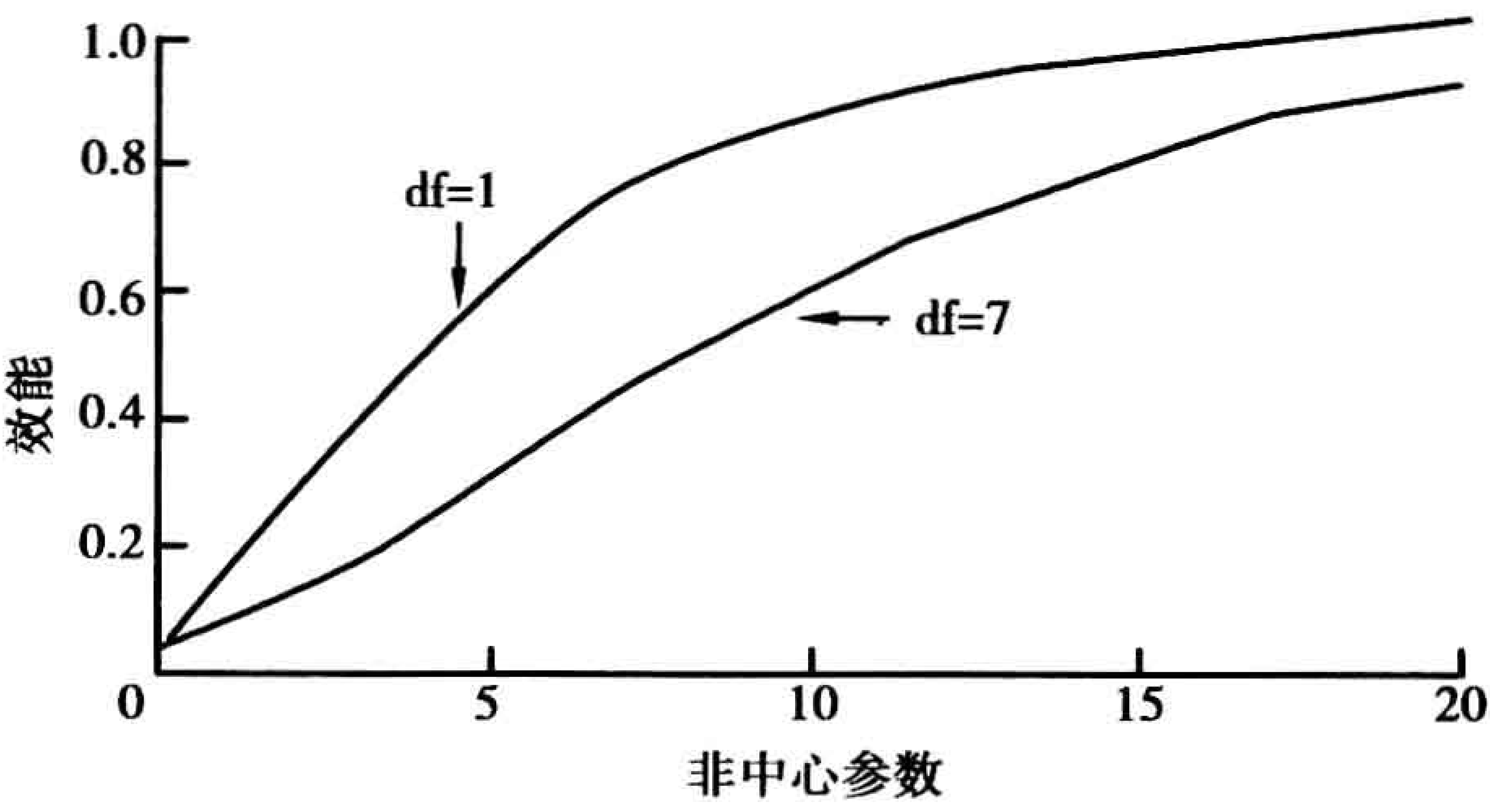


图 6.4 当 df = 1 和 df = 7 时的效能与非中心参数($\alpha = 0.05$)

6.4.4 例子:麻风病的治疗

表 6.11 取自一项使用砜类和链霉素药来治疗麻风病的实验。实验开始时的渗透程度测量的是某种类型的皮肤损害，结果变量是在治疗 48 周后病患临床状况的总体变化。我们对结果变量的赋值为 $\{-1, 0, 1, 2, 3\}$。这里所关注的问题为，对于病情严重程度不同的对象疗效是否也不同。

表 6.11 按照渗透程度划分的临床状况的变化

临床变化	渗透程度		高度渗透的比例
	高	低	
恶化	1	11	0.08
没有变化	13	53	0.20
轻微改善	16	42	0.28
中度改善	15	27	0.36
明显改善	7	11	0.39

来源:授权重印自:the Biometric Society(Cochran 1954)。

在这里，临床变化的结果变量是定序变量。一种自然而然的想法是比较两种渗透水平之间的平均变动。Cochran(1954)和 Yates(1948)指出，这种分析与将二分变量视为结

果变量的趋势检验完全一致。该检验对临床变化和高度渗透案例所占比例之间的线性关系非常敏感。

检验结果 $G^2(I) = 7.28$ (df = 4)并没有给出很明显证据表明存在关联($P = 0.12$),但是该检验忽略了各行间的排序。随着临床状况的改善,高度渗透所占样本比例单调增加。在线性 Logit 模型中,关于 $H_0:\beta=0$的检验结果为 $G^2(I \mid L) = 6.65$, df = 1 ($P = 0.01$)。存在强有力的证据表明,在较高程度的渗透情况下,出现了更积极的临床变化。利用排序信息使检验的自由度从 4 下降为 1,这对检验结果起了很大的作用。另外, $G^2(L) = 0.63$,df = 3,表示线性趋势模型对数据拟合得很好。

6.4.5 通过模型修匀提高估计精度

利用定向的备择假设不仅可以提高*检验效能*(*test power*),也可以改进对单元格概率和综合指标的*估计*(*estimation*)。在一般形式下,记 $\boldsymbol{\pi}$ 为列联表单元格的真实概率,**p** 表示样本比例,令 $\hat{\boldsymbol{\pi}}$ 表示模型对 $\boldsymbol{\pi}$ 的最大似然估计。

当 $\boldsymbol{\pi}$ 满足某个模型时,该模型的 $\hat{\boldsymbol{\pi}}$ 和 **p** 都是关于 $\boldsymbol{\pi}$ 的一致估计。这时,模型估计值 $\hat{\boldsymbol{\pi}}$ 要优于 **p**,因为它的真实的渐近标准误不会大于 **p** 的标准误。这可归因于模型的简约性(parsimony):用于计算 $\hat{\boldsymbol{\pi}}$ 的非饱和模型比计算 **p** 的饱和模型具有更少的参数。事实上,在估计有关单元格概率的函数 $g(\boldsymbol{\pi})$时,模型估计值也更有效。对于任意可导函数 g,

$$\text{var}[\sqrt{n}g(\hat{\pi})] \text{ 的渐近值} \leqslant \text{var}[\sqrt{n}g(p)] \text{ 的渐近值。}$$

有关证明,我们将在第 14.2.2 节给出。这一特性可以推广到分类数据模型以外的更一般的情况(Altham,1984),它是统计学家偏好简约模型的原因之一。

当然,在现实中一个被选中的模型完全成立的可能性不大。不过,当模型能够较好地近似 $\boldsymbol{\pi}$ 时,除了 n 极大的情况外,$\hat{\boldsymbol{\pi}}$ 便优于 **p**。尽管 $\hat{\pi}_i$ 是有偏的,但是它比 p_i 的方差要小,而且当它的方差加上偏差平方项之和仍小于 $\text{var}(p_i)$ 时,存在 $\text{MSE}(\hat{\pi}_i) < \text{MSE}(p_i)$。在第 3.3.7 节我们给出了在二维表中,即使模型不成立时,独立性模型对单元格概率的估计可能仍好于样本比例。

6.5 样本规模与统计效能*

在任何统计分析中,样本规模 n 都会影响结果。即使当 n 很小时,也很可能检测到强烈的效应。相反,对弱效应的研究则需要很大的 n 。研究设计应当考虑在达到特定效能的条件下,考察某一效应所需要的样本规模。

6.5.1 样本规模和比较两个比例的效能

对于服从大样本正态分布的检验统计量,可以通过普通的方法计算效能。我们以比较两种医学干预方式的二项分布参数 π_1 和 π_2 的检验为例。在一项实验中,计划接受每种干预的独立样本的规模均为 $n_i = n/2$ 。研究者期望对于每种干预方式 $\pi_i \approx 0.6$,且两者之间相差 0.10 以上表示重要差异。在检验 $H_0: \pi_1 = \pi_2$ 时,样本间 $\hat{\pi}_1 - \hat{\pi}_2$ 的方差等于 $\pi_1(1-\pi_1)/(n/2) + \pi_2(1-\pi_2)/(n/2) \approx 0.6 \times 0.4 \times (4/n) = 0.96/n$ 。特别地,

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - (\pi_1 - \pi_2)}{(0.96/n)^{1/2}}$$

在 π_1 和 π_2 接近 0.6 时近似服从标准正态分布。

在 α 水平上对 H_0 进行检验的效能近似等于

$$P\left[\frac{|\hat{\pi}_1 - \hat{\pi}_2|}{(0.96/n)^{1/2}} \geqslant z_{\alpha/2}\right]。$$

当 $\pi_1 - \pi_2 = 0.10$ 时,对于 $\alpha = 0.05$;上式等于

$$P\left[\frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0.10}{(0.96/n)^{1/2}} > 1.96 - 0.10(n/0.96)^{1/2}\right] +$$

$$P\left[\frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0.10}{(0.96/n)^{1/2}} < -1.96 - 0.10(n/0.96)^{1/2}\right]$$

$$= P[z > 1.96 - 0.10(n/0.96)^{1/2}] + P[z < -1.96 - 0.10(n/0.96)^{1/2}]$$

$$= 1 - \Phi[1.96 - 0.10(n/0.96)^{1/2}] + \Phi[-1.96 - 0.10(n/0.96)^{1/2}],$$

其中 Φ 表示标准正态累积分布函数。当 $n = 50$ 时,效能大约为0.11;当 $n = 200$ 时,效能约为0.30。如果所检验的效应较小且样本规模也不是很大,很难达到统计显著性。图6.5显示了当 $\pi_1 - \pi_2 = 0.1$ 时效能如何随着 n 的增加而上升。作为对照,图6.5 也给出了当 $\pi_1 - \pi_2 = 0.2$ 时的情况。

对于给定的 P(第一类错误) $= \alpha$ 和 P(第二类错误) $= \beta$(效能 $= 1 - \beta$),我们可以计算为了达到这些值所需要的样本规模。一项同时保证 $n_1 = n_2$ 的研究大约需要

$$n_1 = n_2 = (z_{\alpha/2} + z_\beta)^2[\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)]/(\pi_1 - \pi_2)^2 。$$

当 π_1 和 π_2 的真值分别是0.60 和0.70 时,对于一个要求 $\alpha = 0.05$、$\beta = 0.10$ 的检验,$n_1 = n_2 = 473$。该公式也提供了有关 $\pi_1 - \pi_2$ 的置信区间所需要的样本规模。当每组大约有473 个对象且实际上 $\pi_1 = 0.60$ 和 $\pi_2 = 0.70$ 时,$\pi_1 - \pi_2$ 的95%的置信区间仅有0.10 的可能性包含0。

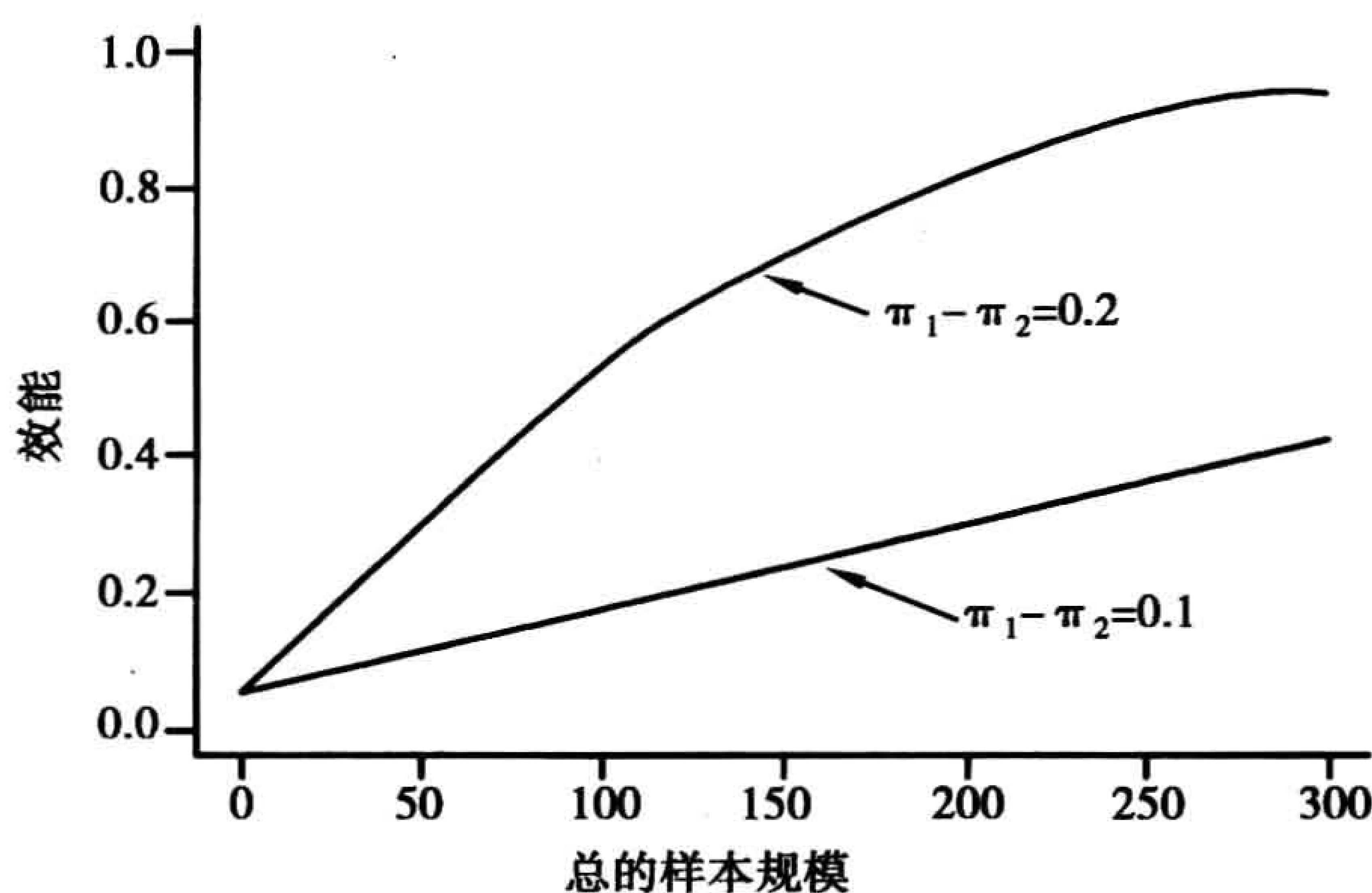


图6.5 关于两个比例相等的统计检验的近似效能,真值接近于取值范围中部且 $\alpha = 0.05$

这个样本规模公式是近似的,而且可能对所要求的实际值略有低估。但在大多数实际应用中,该公式是足够的,其中对 π_1 和 π_2 只能进行粗略的猜测。更精确的公式,参见:Fleiss(1981)。

6.5.2 Logistic 回归中样本规模的决定

现在,考虑模型 Logit $[\pi(x_i)] = \alpha + \gamma x_i$, $i = 1,\cdots,n$,其中 x 为分类变量(我们在这儿使用 γ,是为了避免与 $\beta = P$(第二类错误) 相混淆)。在检验 H_0: $\gamma = 0$时,为了达到一定的效能,模型所需要的样本规模取决于 $\hat{\gamma}$ 的方差。这又取决于 $\{\pi(x_i)\}$,并且关于 n 的公式需要利用对 $\hat{\pi} = \pi(\bar{x})$ 的猜测以及 X 的分布。效应的大小是比较 $\pi(\bar{x})$ 和 $\pi(\bar{x} + s_x)$ 的对数发生比之比 τ,其中 $\pi(\bar{x} + s_x)$ 表示超出 x 的均值一个标准差时的概率。当 X

近似于正态分布时,对于单边检验,Hsieh(1989)推导出

$$n = [z_\alpha + z_\beta \exp(-\tau^2/4)]^2(1 + 2\hat{\pi}\delta)/(\hat{\pi}\tau^2),$$

其中

$$\delta = [1 + (1 + \tau^2)\exp(5\tau^2/4)]/[1 + \exp(-\tau^2/4)]。$$

随着 $\hat{\pi} \to 0.5$ 以及 $|\tau|$ 的上升, n 的值会下降。

我们以在某一人群中 x = 胆固醇水平对患有严重心脏病的概率的效应为例来对此加以说明,在该人群的平均胆固醇水平上,患心脏病的概率大约是0.08。对于胆固醇水平上升一个标准差,研究者期望检验对该概率上升50%具有灵敏度。在胆固醇的均值水平上,严重心脏病的发生比等于 $0.08/0.92 = 0.087$;胆固醇水平比均值高一个标准差所对应的发生比等于 $0.12/0.88 = 0.136$。发生比之比等于 $0.136/0.087 = 1.57$,并且 $\tau = \log(1.57) = 0.450$。这样,对于 $\alpha = 0.05$ 以及 $\beta = 0.10$, $\delta = 1.306$, $n = 612$。

6.5.3 多元 Logistic 回归的样本规模

多元 logistic 回归需要较大的 n 来检测效应。令 R 表示所关注的预测变量 X 与模型中其他变量之间的多元相关系数,则以上关于 n 的公式需除以 $(1 - R^2)$。在该公式中, $\hat{\pi}$ 是在所有的解释变量都取均值时得出的,并且发生比之比指的是在其他预测变量取均值时 X 的效应。

在第6.5.2节的例子中,考虑血压也是一个预测变量。如果胆固醇和血压之间的相关系数等于0.40,我们需要 $n \approx 612/[1 - (0.40)^2] = 729$。

这些公式只是提供了关于样本规模的粗略估计。在大多数应用中,我们只能对 $\hat{\pi}$ 和 R 进行粗略的猜测,而且 X 可能远远不同于正态分布。其他关于这一问题的研究,参见:Hsieh et al. (1998)、Whittemore(1981)。

6.5.4 列联表卡方检验的效能

当假设不成立时,正态分布变量的平方项以及 X^2 和 G^2 统计量服从大样本非中心卡方分布(第6.4.2节)。假定 H_0 等价于关于列联表的模型 M,令 π_i 表示单元格 i 的真实概率, $\pi_i(M)$ 表示模型 M 收敛时的最大似然估计 $\hat{\pi}_i$,其中 $\sum \pi_i = \sum \pi_i(M) = 1$。对于一个规模为 n 的多项分布样本, X^2 的非中心参数等于

$$\lambda = n\sum_i \frac{[\pi_i - \pi_i(M)]^2}{\pi_i(M)}。 \quad (6.8)$$

它具有与 X^2 相同的形式,只是用 π_i 替代了样本比例 p_i,用 $\pi_i(M)$ 替代了 $\hat{\pi}_i$。G^2 的非中心参数等于

$$\lambda = 2n\sum_i \pi_i \log \frac{\pi_i}{\pi_i(M)}。 \quad (6.9)$$

当 H_0 成立时,所有的 $\pi_i = \pi_i(M)$。这时,对于两个统计量,都有 $\lambda = 0$,进而可以适用中心卡方分布。

为了近似确定 $df = v$ 的卡方检验的效能,需要:①选取一组假设的真值 $\{\pi_i\}$,②根据在 H_0 成立时的模型 M 对 $\{\pi_i\}$ 的拟合计算 $\{\pi_i(M)\}$,③计算非中心参数 λ,④计算 $P[X^2_{v,\lambda} > X^2_v(\alpha)]$。表6.12给出了在 $\alpha = 0.05$ 水平上第四步中有关非中心卡方概率表的节选。

表 6.12　在 $\alpha = 0.05$ 水平上卡方检验的效能

	非中心参数													
df	0.0	0.2	0.4	0.6	0.8	1.0	2.0	3.0	4.0	5.0	7.0	10.0	15.0	25.0
1	.050	.073	.097	.121	.146	.170	.293	.410	.516	.609	.754	.885	.972	.998
2	.050	.065	.081	.098	.115	.133	.226	.322	.415	.504	.655	.815	.944	.996
3	.050	.062	.075	.088	.102	.116	.192	.275	.358	.440	.590	.761	.917	.993
4	.050	.060	.071	.082	.093	.106	.172	.244	.320	.396	.540	.716	.891	.989
6	.050	.058	.066	.075	.084	.094	.146	.206	.270	.336	.468	.644	.843	.980
8	.050	.057	.064	.071	.079	.087	.131	.182	.238	.296	.417	.588	.799	.968
10	.050	.056	.062	.068	.075	.082	.121	.166	.215	.268	.379	.542	.760	.956
20	.050	.053	.056	.060	.063	.066	.096	.125	.158	.193	.273	.402	.611	.883
50	.050	.052	.054	.056	.059	.061	.076	.092	.110	.129	.173	.250	.398	.687

来源：授权重印自：G. E. Haynam, Z. Govindarajulu, and F. C. Leone, in *Selected Tables in Mathematical Statistics*, eds. H. L. Harter and D. B. Owen (Chicago: Markham, 1970)。

6.5.5　条件独立性检验的效能

我们使用 O'Brien(1986)中的一个例子。某项标准的胎儿心跳监测检查预测胎儿在出生后是否需要特别看护。标准检查结果的类别为(值得担忧,可靠)。结果变量 Y 指新生儿在产后第一周是否需要某种特别看护(1 = 是,0 = 否)。另一种新的胎儿心跳监测检查被开发出来,其类别包括(非常担忧,比较担忧,可靠)。医师计划研究这种新检查是否能对预测结果有帮助,也即,在给定标准检查结果的情况下,结果变量和新检查结果之间是否存在关联?相应的统计量检验在包括新检查(N)和标准检查(S)作为分类预测变量的 Logit 模型中新监测检查的效应。

为了确定 n,统计学家要求医师猜测解释变量之间的联合分布,比如问"你认为多大比例的对象会在两种检查中都被评为'可靠'?"之类的问题。对于每一种 NS 的取值组合,医师也对 $P(Y = 1)$ 进行了猜测。表 6.13 给出了边际分布和条件概率的一种方案。从它们的乘积可以得到联合分布 $\{\pi_{ijk}\}$,如由标准检查判定为"值得担忧"、新检查判定为"非常值得担忧"且需要特别看护的样本比例为 $0.04 \times 0.40 = 0.016$。这些联合概率给出了空模型以及其他 Logit 模型的拟合概率 $\boldsymbol{\pi}(M_0)$ 和 $\boldsymbol{\pi}(M_1)$(读者可以将 $\{\pi_{ijk}\}$ 的百分比作为计数录入拟合 logistic 回归的软件,拟合相应模型,再将拟合计数除以 100 得出所拟合的联合概率)。比较这些模型的似然比检验具有形式为式 6.9 的非中心参数,其中用 $\boldsymbol{\pi}(M_1)$ 代替 $\boldsymbol{\pi}$,用 $\boldsymbol{\pi}(M_0)$ 代替 $\boldsymbol{\pi}(M)$。

表 6.13　一种计算效能的方案

标准检查	新检查	联合概率	特别看护
值得担忧	非常担忧	0.04	0.40
	比较担忧	0.08	0.32
	可靠	0.04	0.27
可靠	非常担忧	0.02	0.30
	比较担忧	0.18	0.22
	可靠	0.64	0.15

来源：授权重印自：O'Brien(1986)。

对于表 6.13 中的情况,非中心参数等于 $0.00816n$,自由度为 df = 2。对于 n 分别等

于400,600和1 000,当 $\alpha = 0.05$ 时效能大约分别为0.35,0.49和0.73。这种方案预测64%的观测值将发生在预测变量的一个取值组合点。变量离散度的不足削弱了效能。

6.5.6 样本规模对模型选择和统计推断的影响

样本规模的影响意味着,在进行模型选择时需要慎重。在小样本情况下,拟合优度检验所接受的最简约的模型可能会非常简单。相反,大样本通常要求相当复杂的模型才能通过拟合优度检验。所以,有些统计上显著的效应可能很弱而且缺乏实质意义。在大样本的情况下,可能使用一个比通过拟合优度检验更简单的模型就足够了。因此,在分析中只注重拟合优度检验是不够的,估计模型参数以及描述效应的强度也很必要。

以上这些评述仅仅反映了显著性检验的局限性。事实上,零假设很少真正成立。当 n 足够大时,它们总会被拒绝掉。更重要的问题是,参数真值和零假设的值之间的差异是否足够重要。许多方法学家过分强调检验的重要性,而忽视了如置信区间等估计方法。当 P 值较小时,置信区间给出了 H_0 可能不成立的范围,从而帮助我们决定拒绝它是否具有现实意义。当 P 值较大时,置信区间显示出某些可能的参数值是否与 H_0 相差甚远。一个包含了 H_0 值的很大的置信区间表明,对于某些重要的备择假设来说,该检验的统计效能很弱。

6.6 Probit 模型和补余双对数模型*

在本节中,我们讨论除 Logit 模型外的其他两种关于二分结果变量的模型。与 Logit 模型一样,这些模型具有式4.8的形式,

$$\pi(x) = \Phi(\alpha + \beta x), \tag{6.10}$$

其中 Φ 是连续的累积分布函数。下面首先讨论这类模型的渊源。

6.6.1 容差与二分结果变量模型

在毒理学中,二分结果变量模型常用来描述一剂毒素是否会导致研究对象死亡。容差分布(*tolerance distribution*)阐述了式6.10模型的合理性。令 x 表示剂量水平,对于一个随机选中的对象,令 $Y = 1$ 表示该对象死亡。假定研究对象对该剂量的容差为 T,则 $(Y = 1)$ 等价于 $(T \leqslant x)$。例如,一种昆虫在当剂量 x 小于 T 时可以存活,而当剂量 x 大于等于 T 时会死亡。不同对象之间的容差是有差异的,令 $F(t) = P(T \leqslant t)$。对于给定的剂量 x,一个随机选中的对象死亡的概率等于

$$\pi(x) = P(Y = 1 \mid X = x) = P(T \leqslant x) = F(x)。$$

也就是说,适当的二项分布模型应当具有容差分布的累积分布函数 F 的形状。令 Φ 表示 F 所属分布族的标准累积分布函数。利用 T 的均值和方差进行常见的标准化处理,可得

$$\pi(x) = F(x) = \Phi[(x - \mu)/\sigma]。$$

因此,该模型具有 $\pi(x) = \Phi(\alpha + \beta x)$ 的形式。

6.6.2 Probit 模型

毒理学实验通常用浓度的对数来测量剂量水平(Bliss,1935)。一般来说,剂量的容差分布近似于 $N(\mu,\sigma^2)$ 分布,其中 μ 和 σ 为未知参数。如果 F 是 $N(\mu,\sigma^2)$ 的累积分布函数,那么 $\pi(x)$ 具有 $\pi(x) = \Phi(\alpha + \beta x)$ 的形式,其中 Φ 是标准正态累积分布函数,$\alpha = -\mu/\sigma$,$\beta = 1/\sigma$。按照广义线性模型的表达形式,

$$\Phi^{-1}[\pi(x)] = \alpha + \beta x \tag{6.11}$$

就是 *probit 模型*(*probit model*)。Probit 连结函数为 $\Phi^{-1}(\cdot)$。鉴于累积分布函数将实数的取值范围映射到 (0,1) 的概率刻度上,因此,累积分布函数的反函数将 $\pi(x)$ 的 (0,1) 刻度映射到二分结果变量模型中线性预测项的整个实数取值范围上。

$\pi(x)$ (或当 $\beta < 0$ 时, $1 - \pi(x)$)所对应的结果变量曲线具有均值为 $\mu = -\alpha/\beta$ 和标准差为 $\sigma = 1/|\beta|$ 的正态累积分布函数的形状。由于68%的正态分布密度落在距离均值一个标准差的区间内,因而, $1/|\beta|$ 等于在 $\pi(x) = 0.50$ 与 $\pi(x) = 0.16$ 或 0.84 对应的 x 值之间的距离。$\pi(x)$ 的变动率为 $\partial\pi(x)/\partial x = \beta\phi(\alpha + \beta x)$,其中 $\phi(\cdot)$ 为标准正态密度函数。当 $\alpha + \beta x = 0$ 时(即在 $x = -\alpha/\beta$ 时)这个率最大,此时,它等于 $\beta/(2\pi)^{1/2} = 0.40\beta$ (其中 $\pi = 3.14\cdots$)。在这点上, $\pi(x) = \frac{1}{2}$ 。

相比而言,在参数为 β 的 logistic 回归中, $\pi(x)$ 的曲线是一个标准差为 $\pi/|\beta|\sqrt{3}$ 的 logistic 累积分布函数。在 $x = -\alpha/\beta$ 时, $\pi(x)$ 的变动率等于 0.25β 。当 logistic 模型的参数 β 等于 probit 模型中 β 的 0.40/0.25 = 1.6 倍时,probit 和 logistic 曲线的累积分布函数在 $\pi(x) = \frac{1}{2}$ 处对应的变动率相等。当 logistic 的参数 β 等于 probit 参数 β 的 $\pi/\sqrt{3} = 1.8$ 倍时,二者的标准差相等。在两个模型都拟合得很好的情况下,logistic 回归中的参数估计大约是 probit 模型的 1.6 到 1.8 倍。

二项分布回归模型的似然方程公式(式 4.24)也适用于 probit 模型(另见习题6.32)。我们可以利用广义线性模型的 Fisher 计分法来求解似然方程(Bliss, 1935, Fisher, 1935b)。Newton-Raphson 法会给出相同的最大似然估计,但二者的标准误略有差别。对于通过求逆来获取渐近协方差矩阵的信息矩阵,Newton-Raphson 法使用的是观察信息,而 Fisher计分法使用的是期望信息。除 Logit 连结外,它们在其他二项分布连结中并不相同。

6.6.3 例子:甲虫的死亡率

表 6.14 给出了甲虫暴露在不同浓度的二硫化碳气体中 5 小时后被杀死的数量。图 6.6 显示了浓度的对数与相应的甲虫被杀死的比例(图中的点)。在大约 $x = 1.8$ 时,这个比例迅速升高,并在此之后接近于 1。

利用 probit 模型,最大似然拟合结果为

$$\Phi^{-1}[\hat{\pi}(x)] = -34.96 + 19.74x。$$

在该模型中,当 $x = 34.96/19.74 = 1.77$ 时, $\hat{\pi}(x) = 0.50$。它对应于一个 $\mu = 1.77$ 和 $\sigma = 1/19.74 = 0.05$ 的标准容差分布。$\hat{\pi}(x)$ 的曲线对应于 $N(1.77, 0.05^2)$ 的累积分布函数曲线。

在剂量水平 x_i 对应的 n_i 只甲虫中, $n_i\hat{\pi}(x_i)$ 是所拟合的死亡计数, $i = 1,\cdots,8$ 。表 6.14 给出了模型的拟合值,图 6.6 也显示了相应的拟合情况。此外,表 6.14 还给出了线性 Logit 模型的拟合值。这些模型拟合结果相似,但都不是太好。Logit 模型的拟合优度统计量 G^2 等于 11.1,probit 模型的 G^2 等于 10.0,自由度为 df = 6。

表 6.14 二硫化碳所杀死的甲虫数

剂量的对数	甲虫数量	死亡数量	拟合值		
			补余双对数模型	Probit 模型	Logit 模型
1.691	59	6	5.7	3.4	3.5
1.724	60	13	11.3	10.7	9.8
1.755	62	18	20.9	23.4	22.4

续表

剂量的对数	甲虫数量	死亡数量	拟合值		
			补余双对数模型	Probit 模型	Logit 模型
1.784	56	28	30.3	33.8	33.9
1.811	63	52	47.7	49.6	50.0
1.837	59	53	54.2	53.4	53.3
1.861	62	61	61.1	59.7	59.2
1.884	60	60	59.9	59.2	58.8

来源:授权重印自:Bliss(1935)。

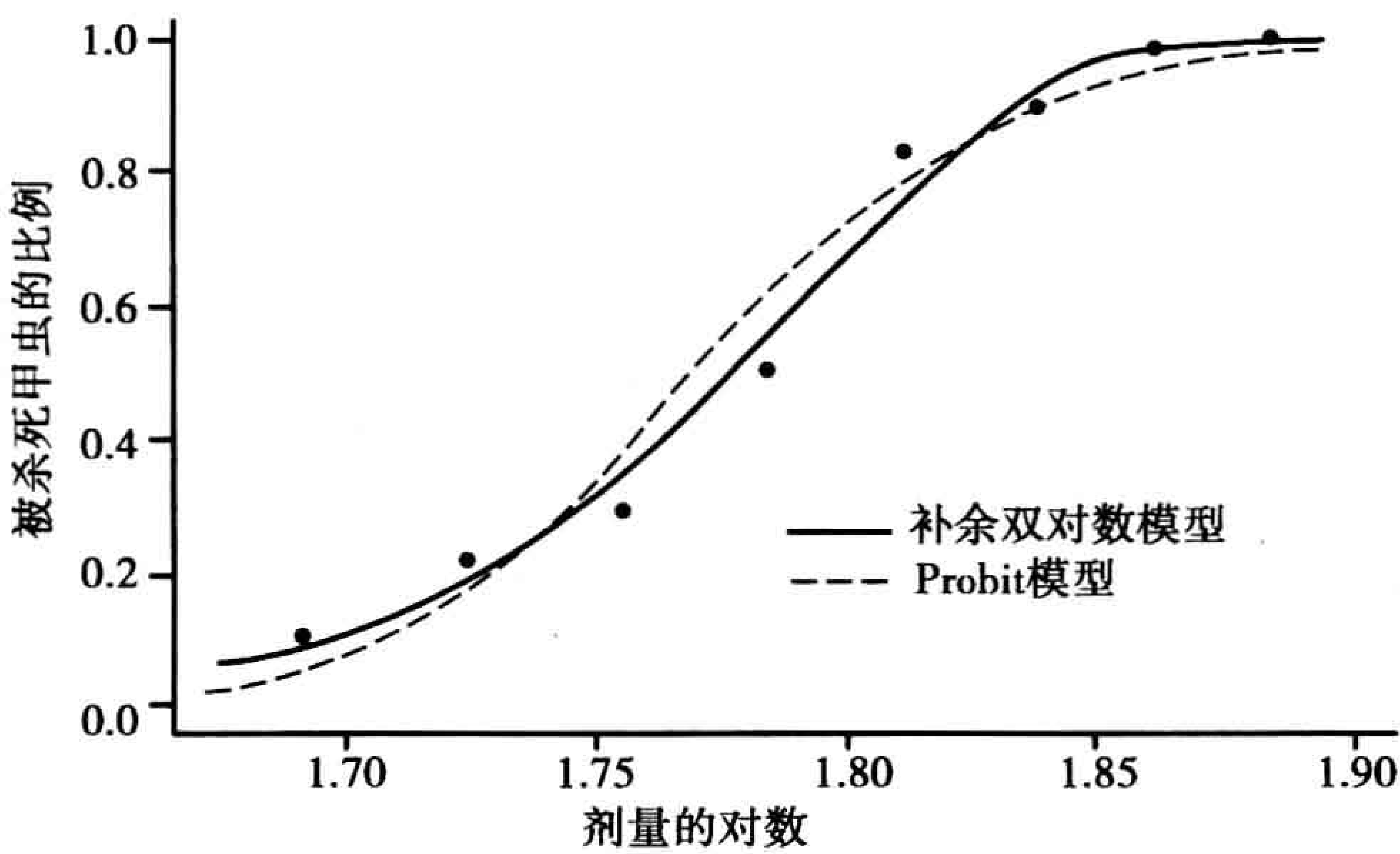


图 6.6　剂量的对数与被杀死甲虫的比例以及 probit 模型和补余双对数模型的拟合结果

6.6.4　补余双对数模型

Logit 和 probit 连结在概率等于 0.5 处是对称的,这意味着:

$$连结[\pi(x)] = -连结[1-\pi(x)]。$$

具体来说,

$$\text{Logit}[\pi(x)] = \log[\pi(x)/(1-\pi(x))]$$
$$= -\log[(1-\pi(x))/\pi(x)] = -\text{Logit}[1-\pi(x)]。$$

这表明,关于 $\pi(x)$ 的结果变量曲线在 $\pi(x) = 0.5$ 处具有对称的形状,所以 $\pi(x)$ 趋近于 0 和它趋近于 1 的速度是相同的。当结果变量曲线存在严重的非对称性时,Logit 和 probit 模型都不适用。

以下结果变量曲线

$$\pi(x) = 1 - \exp[-\exp(\alpha + \beta x)] \tag{6.12}$$

具有图 6.7 所示的形状。该曲线是非对称的,$\pi(x)$ 趋近 0 的速度比它趋近 1 的速度慢。对于这个模型,有:

$$\log[-\log(1-\pi(x))] = \alpha + \beta x。$$

这个广义线性模型的连结函数被称为*补余双对数*(*complementary log-log*)连结,因为它对 $\pi(x)$ 的补函数使用了 log-log 连结。

在解释式 6.12 模型时,注意在点 x_1 和 x_2 处,

$$\log[-\log(1-\pi(x_2))] - \log[-\log(1-\pi(x_1))] = \beta(x_2 - x_1),$$

因此,

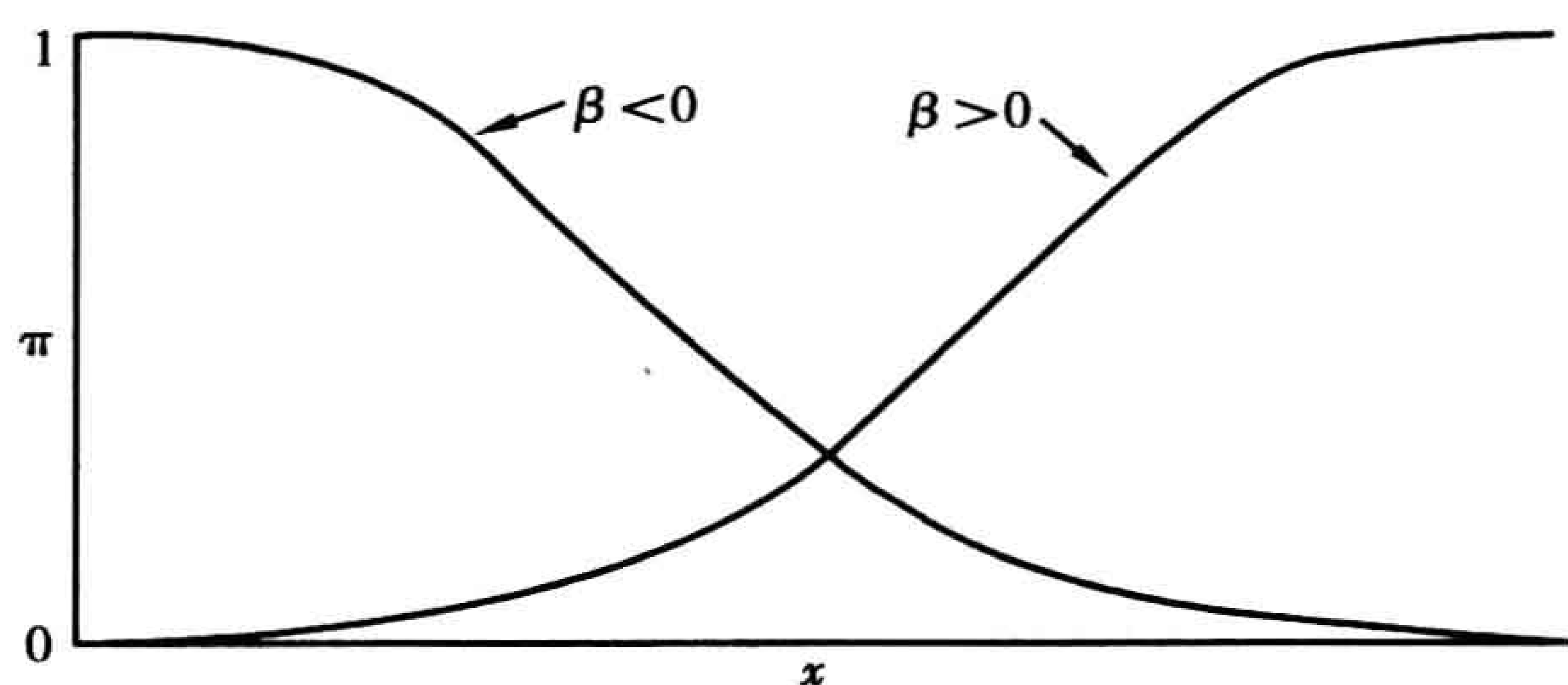


图 6.7 使用补余双对数连结的模型

$$\frac{\log[1-\pi(x_2)]}{\log[1-\pi(x_1)]}=\exp[\beta(x_2-x_1)],$$

并且

$$1-\pi(x_2)=[1-\pi(x_1)]^{\exp[\beta(x_2-x_1)]}。$$

对于 $x_2 - x_1 = 1$，点 x_2 处的补概率等于点 x_1 处补概率的 $\exp(\beta)$ 的乘方。

一个与式 6.12 有关的模型是

$$\pi(x)=\exp[-\exp(\alpha+\beta x)]。\tag{6.13}$$

对于此模型，$\pi(x)$ 迅速趋近于 0 而较缓慢地趋近于 1。随着 x 的上升，当 $\beta>0$ 时该曲线单调递减，当 $\beta<0$ 时它单调递增。按照广义线性模型的形式，它使用的是*双对数*（*log-log*）连结

$$\log[-\log(\pi(x))]=\alpha+\beta x。$$

当补余双对数模型对于某个事件成功的概率成立时，双对数模型则对相应的失败概率成立。

使用双对数连结的式 6.13 模型是*极值或冈布尔*（*extreme value or Gumbel*）分布的累积分布函数（式 6.10）的一个特例。对于参数 $b>0$，$-\infty<a<\infty$，该累积分布函数等于

$$F(x)=\exp\{-\exp[-(x-a)/b]\}。$$

它的均值为 $a\pm0.577b$，标准差为 $\pi b/\sqrt{6}$。使用双对数连结的模型可以通过广义线性模型的 Fisher 计分法来拟合。

6.6.5 例子：再论甲虫的死亡率

对于甲虫死亡率的数据（表 6.14），补余双对数模型的最大似然估计为 $\hat{\alpha}=-39.52$ 和 $\hat{\beta}=22.01$。在剂量 $x=1.7$ 时，所拟合的存活概率为 $1-\hat{\pi}(x)=\exp\{\exp[-39.52+22.01(1.7)]\}=0.885$；而当 $x=1.8$ 时，该概率等于 0.332；当 $x=1.9$ 时，它等于 5×10^{-5}。在剂量 $x+0.1$ 处的存活概率等于剂量 x 处相应概率的 $\exp(22.01\times0.1)=9.03$ 的乘方。例如，$0.332=(0.885)^{9.03}$。

表 6.14 给出了模型的拟合值，同时，图 6.6 显示了相应的拟合情况。这些拟合结果与所观察到的死亡计数很接近（$G^2=3.5$，df = 6），因而该模型的拟合看上去是充分的。对这些数据的进一步讨论，参见：Aranda-Ordaz（1981）、Stukel（1988）。

6.7 条件 Logistic 回归与精确分布*

当样本规模 n 比参数数量大很多时，logistic 模型参数的最大似然估计表现最好。当 n 较小或者参数数量随着 n 的增加而增加时，利用*条件最大似然法*（*conditional maximum*

likelihood)可以更好地进行统计推断。在这一节,我们介绍条件似然法,并将在第 10.2 节利用该方法分析匹配的个案-控制研究。

6.7.1 条件似然法

条件似然法通过以充分统计量为条件,剔除了冗余参数。它是对 2 × 2 表格的 Fisher 精确法(第 3.5 节)的一般化。条件似然法对应于一种对潜在样本所界定的条件分布,这种潜在样本能够提供与观察样本中出现的冗余参数同样的信息。

我们首先阐述一般性的情况,然后再讨论一些特例。令 y_i 表示第 i 个对象的二分结果变量的值, $i = 1,\cdots,N$ (从现在起,每个 y_i 指一次单一试验,所以 $n_i = 1$)。令 x_{ij} 表示该对象对应的预测变量 j 的取值, $j = 1,\cdots,p$ 。这个模型可表示为

$$P(Y_i = y_i) = \frac{\exp[y_i(\alpha + \sum_{j=1}^{p}\beta_j x_{ij})]}{1 + \exp[\alpha + \sum_{j=1}^{p}\beta_j x_{ij}]}, \tag{6.14}$$

上式中代入 $y_i = 1$,便可得通常的表述,如式 5.15。这里,我们将截距和 p 个预测变量的系数明确区分开来。对于 N 个独立的观测值,

$$P(Y_1 = y_1,\cdots,Y_N = y_N) = \frac{\exp[(\sum_i y_i)\alpha + \sum_{j=1}^{p}(\sum_i y_i x_{ij})\beta_j]}{\prod_i[1 + \exp(\alpha + \sum_{j=1}^{p}\beta_j x_{ij})]}。 \tag{6.15}$$

由这个似然函数可知, β_j 的充分统计量为 $\sum_i y_i x_{ij}$, $j = 1,\cdots,p$, α 的充分统计量为 $\sum_i y_i$,即成功发生的总数。

通常来说,模型中总有一些参数是研究所关注的重点。其他的参数可能是为了调整相应效应的需要而包括进来,它们的值并没有什么特别的意义。我们可以以它们的充分统计量为条件,从似然方程中消除这些参数。下面,我们以消除 α 为例来加以说明(在第 10.2.5 节所介绍的有关匹配的个案-控制研究的模型中,在对核心参数进行统计推断时,截距项的存在会导致问题,这时,将其消除掉是必要的)。由于 α 的充分统计量为 $\sum_i y_i$,我们以 $\sum_i y_i$ 为条件,假定 $\sum_i y_i = t$,将具有与 $\sum_i y_i$ 的观察值相同的样本的条件参照集表示为

$$S(t) = \{(y_1^*,\cdots,y_N^*):\sum_i y_i^* = t\}。$$

对于满足 $\sum_i y_i = t$ 的 $\{y_i\}$,条件似然方程等于

$$P(Y_1 = y_1,\cdots,Y_N = y_N \Big| \sum_i y_i = t) = \frac{P(Y_1 = y_1,\cdots,Y_N = y_N)}{\sum_{S(t)}P(Y_1 = y_1^*,\cdots,Y_N = y_N^*)}$$

$$= \frac{\exp[t\alpha + \sum_{j=1}^{p}(\sum_i y_i x_{ij})\beta_j]\Big/\prod_i[1 + \exp(\alpha + \sum_{j=1}^{p}\beta_i x_{ij})]}{\sum_{S(t)}\exp[t\alpha + \sum_{j=1}^{p}(\sum_i y_i^* x_{ij})\beta_j]\Big/\prod_i[1 + \exp(\alpha + \sum_{j=1}^{p}\beta_j x_{ij})]}$$

$$= \frac{\exp[\sum_{j=1}^{p}(\sum_i y_i x_{ij})\beta_j]}{\sum_{S(t)}\exp[\sum_{j=1}^{p}(\sum_i y_i^* x_{ij})\beta_j]}。$$

该方程不受 α 的影响。

条件似然法的使用与普通似然法相同,即对于似然函数中所包括的参数,它们的条件最大似然估计等于使函数最大化的值。通过迭代法来计算,这些估计值渐近服从正态分布,其协方差矩阵等于条件对数似然函数的二阶偏导数矩阵的负逆矩阵。

6.7.2 Logistic 回归的小样本条件推断

在小样本的情况下，可以利用在消除掉所有其他参数后的条件分布对有关参数进行推断。通过这个条件分布，可以计算某个精确的而不是粗略近似的概率，如 P 值（Cox，1970）。

例如，假定我们关注的是对式6.14模型中 β_p 的推断。为了消除其他参数，我们以这些参数的充分统计量 $T_j = \sum_i y_i x_{ij}$ 为条件，$j = 0,\cdots,p-1$（其中 $x_{i0} = 1$）。由上文的讨论，可以得出以下条件分布：

$$P(Y_1 = y_1,\cdots,Y_N = y_N \mid T_j = t_j, j = 0,\cdots,p-1)$$

$$= \frac{\exp[(\sum_i y_i x_{ip})\beta_p]}{\sum_{S(t_0,\cdots,t_{p-1})}\exp[(\sum_i y_i^* x_{ip})\beta_p]} = \frac{\exp(t_p\beta_p)}{\sum_{S(t_0,\cdots,t_{p-1})}\exp(t_p^*\beta_p)},$$

其中

$$S(t_0,\cdots,t_{p-1}) = \{(y_1^*,\cdots,y_N^*) : \sum_i y_i^* x_{ij} = t_j,\quad j = 0,\cdots,p-1\}。$$

该条件分布仅取决于 β_p 。在给定其他参数的情况下，关于 β_p 的推断利用了它的充分统计量 $T_p = \sum_i y_i x_{ip}$ 的条件分布。令 $c(t_0,\cdots,t_{p-1},t)$ 表示在 $S(t_0,\cdots,t_{p-1})$ 中满足 $T_p = t$ 的数据向量的个数。T_p 的条件分布是

$$P(T_p = t \mid T_j = t_j, j = 0,\cdots,p-1) = \frac{c(t_0,\cdots,t_{p-1},t)\exp(t\beta_p)}{\sum_u c(t_0,\cdots,t_{p-1},u)\exp(u\beta_p)}, \quad (6.16)$$

其中分母是对 T_p 的可能值 u 求和。

当检验 $H_0: \beta_p = 0$时，条件分布也可以简化。对于 $H_a: \beta_p > 0$以及观察到的 $T_p = t_{观察值}$，精确的条件 P 值等于

$$\sum_{t \geq t_{观察值}} P(T_p = t \mid T_j = t_j, j = 0,\cdots,p-1) = \frac{\sum_{t \geq t_{观察值}} c(t_0,\cdots,t_{p-1},t)}{\sum_u c(t_0,\cdots,t_{p-1},u)},$$

即，在条件集里 β_p 的充分统计量至少与观察值一样大的数据结构出现的比例。在进行这样的推断时，需要计算 $\{c(t_0,\cdots,t_{p-1},u)\}$ 。除非是最简单的情况，这种计算强度往往很大，而且需要专门的软件（如 CytelSoftware 的 LogXact 或者 SAS 的 PROC LOGISTIC）。在本节剩下的内容中，我们考虑一些小样本推断的特例。

6.7.3 关于 2 × 2 列联表的小样本条件推断

首先，考虑只有一个预测变量 x 的 logistic 回归，

$$\text{Logit}[P(Y_i = 1)] = \alpha + \beta x_i,\quad i = 1,\cdots,N, \quad (6.17)$$

这里 x_i 只取两个值。该模型适用于2×2表格，其中 $x_i = 1$ 表示第1行，$x_i = 0$ 表示第2行。α 的充分统计量为 $\sum_i y_i$，也即第一列的总计。β 的充分统计量为 $T = \sum_i y_i x_i$，在此可简化为第一行中成功发生的次数。等价地，模型的充分统计量就是两行中成功发生的次数。令 s_1 和 s_2 表示这两个二项分布变量，行总计 n_1 和 n_2 是它们的基数（indices）。

为了消除 α，我们以 $s = s_1 + s_2$，即第一列的总计为条件。由于 $N = n_1 + n_2$ 是给定的，因而，这时另一列的边际总计也是给定的。表格的两个边际总计都给定的情况下，得到的是关于 s_1 的超几何概率，它仅取决于 β（参见式3.20，其中令 $\theta = \exp(\beta)$）。这种情

况下，条件分布满足式 6.16，其中 $c(t_0,t) = \binom{n_1}{t}\binom{N-n_1}{t_0-t}$，$t_0 = s$，$t = s_1$。由此得到的关于 $\beta = 0$ 的精确条件检验就是 2 × 2 表格的 Fisher 精确检验（第 3.5.1 节）。

6.7.4 线性 Logit 模型的小样本条件推断

线性 Logit 模型 $\text{Logit}(\pi_i) = \alpha + \beta x_i$，适用于行变量为定序变量的 $I \times 2$ 表格。我们在第 5.3.4 节介绍了这个模型。在该模型中，数据 $\{y_i\}$ 是 I 个独立的 $\{\text{bin}(n_i,\pi_i)\}$ 计数，其中行总计 $\{n_i\}$ 是给定的。以 $s = \sum y_i$ 为条件，列总计可表示为一个与 α 无关的条件似然函数。关于 β 的精确推断利用的是它的充分统计量 $T = \sum x_i y_i$。根据式 6.16，它的分布具有以下形式：

$$P\left(T = t \mid \sum_i y_i = s;\beta\right) = \frac{c(s,t)e^{\beta t}}{\sum_u c(s,u)e^{\beta u}}。\tag{6.18}$$

这里，$c(s,u)$ 等于在所有给定边际总计满足 $T = u$ 的表格中对 $\left[\prod_i \binom{n_i}{y_i}\right]$ 的求和。

当 $\beta = 0$ 时，单元格计数服从多元超几何分布（式 3.19）。为了对此进行检验，将具有给定边际的表格按照 T 进行排序等价于将它们按照 Cochran-Armitage 统计量排序（第 5.3.5 节）。因而，这个关于线性 Logit 模型的检验是一个精确趋势检验。

在第 5.3.5 节中，我们利用 Cochran-Armitage 检验分析了表 5.3 中关于母亲饮酒与婴儿畸形的数据。尽管 $n = 32\ 573$，但是这个表格的分布极不均衡，其中有的计数非常小，有的非常大。因而，对该数据使用小样本方法更可靠。在使用相同赋值的精确条件趋势检验中，关于 $H_a: \beta > 0$ 的单边 P 值等于 0.016 8。双边检验的 P 值等于 0.017 2，这反映了在给定边际计数情况下条件分布的非对称性。这个结果与大样本 Cochran-Armitage双边检验所得到的 P 值 0.010 差别不大。

6.7.5 2 × 2 × K 表格的小样本条件独立性检验

对于 $2 \times 2 \times K$ 表格 $\{n_{ijk}\}$，Cochran-Mantel-Haenszel 检验使用的是 $\sum_k n_{11k}$。就 Logit 式 6.4 模型而言，这是 β（即 X 的效应）的充分统计量。为了对 $\beta = 0$ 进行小样本检验，需要消除其他的模型参数。似然函数的构建过程表明，$\{\beta_k^Z\}$ 的充分统计量是每个分表的列边际总计 $\{n_{+jk}\}$。当 X 和 Z 都是预测变量时，将 XZ 的每个取值组合处的试验次数 $\{n_{i+k}\}$ 视为给定的。因此，关于 β 的精确推断以每层的行总计和列总计为条件。

以层的边际为条件，精确检验利用了 $\sum_k n_{11k}$。在每个分表中，$\{n_{11k}, k = 1,\cdots,K\}$ 的独立零分布以超几何概率发生。K 个密度函数的乘积给出了 $\{n_{11k}, k = 1,\cdots,K\}$ 的联合零分布（这对应于下面的公式 6.19，其中设定 $\theta = 1$），它决定了 $\sum_k n_{11k}$ 的零分布。对于 $H_a: \beta > 0$，P 值为在给定层边际总计的情况下 $\sum_k n_{11k}$ 至少不小于观察值的零分布概率。对此，Mehta 等（1985）介绍了一种快速算法。当 $K = 1$ 时，该检验简化为Fisher精确检验。

6.7.6 例子：升职歧视问题

表 6.15 所示为具备相似资历的美国政府计算机专家获得晋升情况的数据。该表给

出了按照雇员种族划分的三个不同月份的晋升情况。我们检验升职决定与种族的条件独立性,即在式6.4模型中 $H_0: \beta = 0$ 。该表中有几个计数很小,尽管总的样本规模还可以($n = 74$),但是其中一个边际计数(将几个月的数据合并)等于零,所以对此数据使用CMH 检验可能更明智。

对于 $H_a: \beta < 0$ (即发生比之比 < 1),即黑人雇员获得晋升的概率低于白人雇员。就表6.15 的分表的边际分布来看,n_{111} 的取值范围为0 到4,n_{112} 的取值范围为0 到4,而 n_{113} 的取值范围为0 到2。总计 $\sum_k n_{11k}$ 可能的取值范围为0 到10。样本数据是在所有可能情况中的最极端情况,即观察值为 $\sum_k n_{11k} = 0$,检验 P 值就是这个结果发生的零分布概率。软件给出 $P = 0.026$ 。对所有比观察到的表格更不可能发生的表格的概率求和可得,双边检验的 P 值等于0.056。

表6.15 按照种族和月份划分的升职情况

种族	七月的升职情况		八月的升职情况		九月的升职情况	
	是	否	是	否	是	否
黑人	0	7	0	7	0	8
白人	4	16	4	13	2	13

来源:J. Gastwirth, *Statistical Reasoning in Law and Public Policy*(San Diego, CA: Academic Press, 1988), p. 266。

6.7.7 精确条件估计与发生比之比的比较

对于 $2 \times 2 \times K$ 表格的同质性关联模型(式6.4),发生比之比 $\theta = \exp(\beta)$ 的普通最大似然估计值在稀疏数据渐近特性下表现很差。条件最大似然估计值以其他参数的充分统计量为条件,在减小参数空间的基础上再最大化条件似然函数(Anderson, 1970; Birch, 1964b)。

就单元格计数 $\{n_{ijk}\}$ 而言,对于所有的 k 给定 $\{n_{i+k}, n_{+jk}\}$,$(n_{111} = t_1, \cdots, n_{11K} = t_K)$ 的条件概率密度函数是每个层的函数(式3.20)的乘积,即

$$\prod_k P(n_{11k} = t_k \mid n_{1+k}, n_{+1k}, n_{++k}; \theta) = \prod_k \frac{\binom{n_{1+k}}{t_k}\binom{n_{++k} - n_{1+k}}{n_{+1k} - t_k}\theta^{t_k}}{\sum_u \binom{n_{1+k}}{u}\binom{n_{++k} - n_{1+k}}{n_{+1k} - u}\theta^u}。 \quad (6.19)$$

条件最大似然估计值 $\hat{\theta}$ 使式6.19 最大化。与 Mantel-Haenszel 估计值 $\hat{\theta}_{MH}$ 一样,它在标准渐近情况和稀疏数据渐近情况下都具有很好的特性(Anderson, 1970; Breslow, 1981)。由于参数的数量不会随着 K 的变动而变动,它通常比 $\hat{\theta}_{MH}$ 的效力更高一些;当 $\theta = 1.0$ 时,二者的效力相同;在配对数据的情况下二者完全一致(Breslow, 1981)。

对于 $\sum_k n_{11k}$,式6.19 条件分布趋近于1,它被用于检验 $H_0: \theta = \theta_0$,其中 θ_0 可以是任意值。这时,关于 θ 的95%的置信区间包括 P 值超过0.05 的所有 θ_0 。这样的区间保证至少具有与名义水平一样大的涵盖概率(Gart, 1970; Kim and Agresti, 1995; Mehta et al., 1985)。它是对单个 2×2 表格置信区间的扩展(第3.6.1 节)。就升职歧视的例子而言(表6.15),$\sum_k n_{11k} = 0$,所以任何关于 θ 的置信区间的下界都应当是0。将 Cornfield 尾部法区间应用于多个层的情况,StatXact 所给出的95%的置信区间为 $(0, 1.01)$ 。

Zelen(1971)介绍了一个关于发生比之比同质性的小样本检验。对此方法以及有关

列联表的其他小样本方法的讨论,参见:Agresti(1992)。

6.7.8 例子:痢疾的数据

最后,我们考虑一个具有多个变量的例子。表6.16所示为在一家医院住院的2 493名病人的情况。结果变量是他们在住院期间是否患过急性痢疾。三个预测变量包括年龄(超过50岁为1,不足50岁为0),住院时长(超过1周为1,不足1周为0),以及是否使用了头孢氨苄(Cephalexin)抗菌素(是为1,否为0)。我们通过一个只包括主效应项的模型来估计在控制了年龄和住院时长后使用头孢氨苄的效应。

表6.16 精确条件 Logistic 回归的例子

头孢氨苄[a]	年龄[a]	住院时长[a]	痢疾发生数	样本规模
0	0	0	0	385
0	0	1	5	233
0	1	0	3	789
0	1	1	47	1 081
1	1	1	5	5

a 关于0和1的定义,参见正文。

来源:基于康奈尔医学中心 E. Jaffe 和 V. Chang 的研究,详见 LogXact 用户手册(Cambridge, MA: CYTEL Software, 1999),第259页。

该数据的样本规模很大,但其中只有很少的患者得了急性痢疾。另外,所有使用过头孢氨苄的病人都出现了急性痢疾。这种结果变量的所有取值都落在一个类别的边界性结果会导致对某些模型参数的最大似然估计为无穷大。关于头孢氨苄效应的最大似然估计为 ∞ 意味着,随着对头孢氨苄的参数估计上升到无穷大,似然函数持续上升。

为了探讨头孢氨苄的效应,我们利用一个以其他预测变量的充分统计量为条件的精确分布。尽管对头孢氨苄效应的对数发生比之比的参数估计是无穷大,仍然有可能利用条件分布、通过一组对该参数检验的逆转换来构建置信区间。这种方法给出的关于发生比之比的95%的置信区间为(19,∞)。假定主效应模型成立,看上去头孢氨苄具有很强的效应。类似地,关于对数发生比之比等于零的假设检验的结果为 $P<0.000\ 1$。

因为在年龄和住院时长取值组合的前三个点都没有出现使用头孢氨苄的案例,我们对这一结果持保留态度。事实上,表6.16中的前三行对这里的分析没有任何贡献(习题6.18)。该数据关于头孢氨苄效应的结论只适用于住院时间较长的老年患者。

6.7.9 离散性带来的问题

与 Fisher 精确检验一样,关于列联表的精确条件推断也会因离散性问题而偏于保守。当 n 很小或数据不均衡时,即大多数观测值落在某一列或某一行,这个问题尤为突出。使用中位 P 值或是基于对检验和相应置信区间的样本空间进行更细致分割的 P 值(注解3.9)可以缓解保守性的问题。以升职歧视的数据(表6.15)为例,我们给出的关于共同发生比之比的95%的置信区间为(0,1.01)。利用中位 P 值并对精确检验 $H_0: \theta = \theta_0$ 进行逆转换,所得出的区间为(0,0.78)。不过,这种方法并不能保证实际的涵盖概率不小于0.95。

当不存在另一组 $\{y_i^*\}$ 具有与观察数据的某个给定充分统计量 $\sum_i y_i x_{ij}$ 相同的取值时,会出现一个特别的问题。这种情况下,核心参数的充分统计量的条件分布是一种退化(degenerate)分布。因而,精确检验的 P 值等于1.0。这种情况一般发生在至少一个需

要控制的解释变量是连续的并具有不等距的取值时。

最后,条件方法的局限性之一是需要冗余参数的充分统计量。只有在使用典型连结的广义线性模型中,这些充分统计量才存在。因此,条件法可以应用于诸如 Logit 模型,但对 probit 模型却不适用。

注 解

第 6.1 节:模型选择的策略

6.1 与 AIC 类似的模型选择准则是根据贝叶斯理论提出的贝叶斯信息准则 BIC = [G^2 − (log n)(df)]。BIC 考虑了样本规模的影响。与 AIC 相比,随着 n 的增加,BIC 对较复杂模型的趋近更慢。有关细节及评论,参见:Raftery(1986)、the February 1999 issue of *Sociological Methods and Research*。

6.2 树状模型(如 CART)是除 logistic 回归外的另一种可选方法,它通过一系列有次序问题来表示决策过程,根据对象对这些问题的回答生成不同方向的分支。应用 CART 的一个例子是确定某一对象的胸痛是否是由于心脏病发作引起的。Zhang 等(1998)回顾了这种方法。

第 6.2 节:Logistic 回归诊断

6.3 关于 logistic 回归的诊断,参见:Copas(1988)、Fowlkes(1987)、Hosmer and Lemeshow(2000,chap. 5)、Johnson(1985)、Landwehr et al.(1984)、Pregibon(1981)。分别进行诊断有助于检查广义线性模型的各个组成部分的充分性(McCullagh and Nelder, 1989,chap. 12)。对于参数为 γ 的一组连结函数 $g(\mu;\gamma)$,Pregibon(1980)给出了如何估计 γ 以得到使拟合结果最优的连结方法,以及如何检查某个给定连结 $g(\mu;\gamma_0)$ 的充分性。

6.4 Amemiya(1981)、Efron(1978)、Maddala(1983)、Zhang 和 Agresti(2000)以及文中所引述的文献回顾了二分结果变量回归的 R^2 指标。Hosmer 和 Lemeshow(2000,sec. 5.2.3)讨论了分类表及其局限性。Pepe(2000)及其所引文献回顾了 ROC 方法。

第 6.3 节:2 × 2 × K 表格中条件关联的统计推断

6.5 诸如 $\hat{\theta}_{MH}$ 等综合指标同时描述多个层的比例之差或相对风险(Greenland and Robins,1985)。Breslow 和 Day(1980,p. 142)提出了另一种关于发生比之比同质性的大样本检验。在每个分表中,令 $\{\hat{\mu}_{ijk}\}$ 具有与观察数据相同的边际,但令其发生比之比等于 $\hat{\theta}_{MH}$。他们的检验统计量是比较 $\{n_{ijk}\}$ 和 $\{\hat{\mu}_{ijk}\}$ 的皮尔逊统计量。Tarone(1985)表明,由于 $\hat{\theta}_{MH}$ 的效力较低,必须对 Breslow-Day 统计量进行调整才能使其极限服从 df = K − 1 的卡方零分布。这种调整通常都很小。Jones 等(1989)回顾了几个同质性检验并在稀疏和非稀疏数据情况下进行了对比。有关比较发生比之比以及估计其共同值的其他研究包括:Breslow and Day(1980,sec. 4.4)、Donner and Hauck(1986)、Liang and Self(1985)。关于对发生比之比的模型分析,参见:Breslow(1976)、Breslow and Day(1980,sec. 7.5)、Prentice(1976a)。Breslow 重点讨论了回顾性研究,这时,由于结果的总计是给定的,使用条件法是一种自然的选择。

第 6.5 节:样本规模与统计效能

6.6 关于在比较比例时如何确定样本规模,Fleiss(1981,sec. 3.2)给出了相应的表格。对于 $I \times J$ 表的情况,参见:Lachin(1977)。推导了卡方统计量的渐近非零分布特性有:Chapman and Meng(1966)、Drost et al.(1989)、Haberman(1974a,pp. 109-112)、Harkness and Katz(1964)、Mitra(1958)、Patnaik(1949)。另见第 14.3.5 节。O'Brien

(1986)的模拟结果表明,对 G^2 的非中心卡方近似在效能的大多取值范围内都成立。Read 和 Cressie(1988,pp. 147-148)列出了关于 X^2 和 G^2 的非零分布特性的其他研究。

第 6.6 节:Probit 模型和补余双对数模型

6.7 有关 probit 模型的经典文献见:Finney(1971)。Chambers 和 Cox(1967)表明,除非 n 极大,很难区分 probit 和 Logit 模型。Ashford 和 Sowden(1970)将 probit 模型扩展到多元二分结果变量的情况;另见:Lesaffre and Molenberghs(1991)、Ochi and Prentice(1984)。Wedderburn(1976)证明,probit 和补余双对数连结的对数似然函数均是凹函数。

第 6.7 节:条件 Logistic 回归

6.8 有关条件 logistic 回归的详细讨论见第 10.2 节,另参见:Breslow and Day(1980, chap. 7)、Cox(1970)、Hosmer and Lemeshow(2000, chap. 5)。Liang(1984)证明,在指数分布族样本的情况下,条件最大似然估计值和条件计分检验渐近等价于各自相应的非条件情况。有关利用条件似然函数进行精确推断的讨论,参见:Hirji et al.(1987)、Mehta and Patel(1995)、LogXact manual(Cytel Software)。Mehta 等(2000)讨论了相应的蒙特卡洛近似。

习题

应用部分

6.1 对马蹄蟹数据拟合一个以体重和壳宽为预测变量的模型:(a)对 $H_0: \beta_1 = \beta_2 = 0$ 进行似然比检验,(b)分别对偏效应进行检验。为什么(b)部分的两个检验都不显著,而(a)部分的检验却非常显著?

6.2 参考习题 8.13 的数据,将对婚前性行为的态度作为结果变量,通过后向剔除法来选取模型,并加以解释。

6.3 参见表 6.4,拟合由 $(E*P+G)$ 所表示的第三阶段模型。通过参数估计来解释 G 的效应以及 E 对 P 的效应依赖。

6.4 讨论表 6.7 中出现辛普森悖论的原因。

6.5 参考习题 2.12。

a. 拟合一个包括 C 和 D 的主效应的模型。利用它估计 AG 的条件发生比之比。与边际发生比之比相比,说明为什么二者不相等。检验模型的拟合优度。

b. 给定系别,拟合不包括 G 的模型。利用 X^2 检验模型的拟合情况,求出残差,并解释拟合不足的原因(每个系都具有一个单独的非冗余标准化皮尔逊残差。这些残差满足 $\sum_{i=1}^{6} r_i^2 = X^2$,它们的平方项是 X^2 的 6 个 df = 1 的组成部分)。

c. 去除系别 A 后,重新拟合上述两个模型,再次考察是否存在拟合不足,并加以解释。

6.6 对表 6.11 的独立性模型进行残差分析,结果表明存在哪种类型的拟合不足?

6.7 表 6.17 所示为在注射了致命 β 溶血链球菌的兔子中,立即注射盘尼西林和推迟 $1\frac{1}{2}$ 小时后再注射的效果。

a. 令 X = 是否推迟,Y = 是否治愈,Z = 盘尼西林的水平,拟合式 6.4 Logit 模型,阐述单元格出现零计数的模式表明(在不包括截距项的模型中)$\hat{\beta}_1^Z = -\infty$ 和 $\hat{\beta}_5^Z$ =

∞。统计软件给出的结果是什么？

b. 利用这个 Logit 模型，构建关于 *XY* 条件独立性的似然比检验，加以解释。

表 6.17　习题 6.7 的数据

盘尼西林水平	是否推迟	结果	
		治愈	死亡
$\frac{1}{8}$	无	0	6
	$1\frac{1}{2}$ h	0	5
$\frac{1}{4}$	无	3	3
	$1\frac{1}{2}$ h	0	6
$\frac{1}{2}$	无	6	0
	$1\frac{1}{2}$ h	2	4
1	无	5	1
	$1\frac{1}{2}$ h	6	0
4	无	2	0
	$1\frac{1}{2}$ h	5	0

来源：授权重印自：Mantel(1963)。

c. 利用 Cochran-Mantel-Haenszel 检验来检验 *XY* 的条件独立性，加以解释。

d. 利用(i) Logit 模型的最大似然估计，(ii) Mantel-Haenszel 估计，来估计 *XY* 条件发生比之比，加以解释。

e. 单元格计数很小使得大样本分析并不可靠。构建相应的小样本推断，并加以解释。

6.8　参见表 2.6。利用 CMH 统计量来检验在控制了被告人种族后，死刑判罚与受害人种族之间的独立性。给出对该假设的另一种检验，并将两种检验的结果进行对比。

6.9　在按照有关协变量进行匹配的 40 对研究对象中，将干预 A 和干预 B 对一个二分结果变量的影响加以比较。就每一对研究对象而言，干预方式的分配是随机的。对于每种干预方式，有 20 对研究对象结果相同，有 6 对接受干预 A 的结果为成功而干预 B 的结果为失败，剩下的 14 对接受干预 B 的结果为成功而干预 A 的结果为失败。利用 Cochran-Mantel-Haenszel 方法检验结果变量与干预方式的独立性(在第 10.1 节中，我们将介绍一种等价的检验，即 McNemar 检验。)

6.10　参见第 6.5.1 节。假定 $\pi_1 = 0.7$，$\pi_2 = 0.6$，当 $\alpha = 0.05$ 时，需要多大的样本规模以保证下列检验具有大约 0.80 的效能：(a) $H_a: \pi_1 \neq \pi_2$，(b) $H_a: \pi_1 > \pi_2$？

6.11　参见第 6.5.1 节。假定 $\pi_1 = 0.63$，$\pi_2 = 0.57$，当不同干预的样本规模相等时，说明为什么 2×2 表格中的联合概率在干预 A 对应的行中为 0.315 和 0.185，而在干预 B 对应的行中为 0.285 和 0.215。对于独立性模型，说明为什么每行中所拟

合的成功的联合概率为 0.30 而失败的概率为 0.20。证明 X^2 的非中心参数等于 0.003 75n，并且 df = 1。当 n = 200 且 α = 0.05 时，求相应的效能。

6.12 在某项对两种干预方式进行比较的实验中，结果变量具有三个类别，研究者期望的条件分布大约分别为 (0.2,0.2,0.6) 和 (0.3,0.3,0.4)。

a. 在 α = 0.05 的水平上，利用(i) X^2，(ii) G^2，求在每个干预组都有 100 个观测值时比较两个分布的效能。比较相应结果。

b. 要使(a)部分中的检验具有 0.90 的效能，每个干预组大约需要多大的样本规模？

6.13 表 4.3 中马蹄蟹壳宽的值具有 $\bar{x}$ = 26.3 和 s_x = 2.1。如果真实的关系与第 5.1.3 节中的拟合方程相似，对于 $H_0:\beta=0$ 和 $H_a:\beta>0$ 的检验，大约需要多大的样本规模可使得在 α = 0.05 的水平上 P(第二类错误) = 0.10？

6.14 参见习题 5.1。表 6.18 给出了一个 probit 模型的拟合结果。通过以下方式对参数估计进行解释：(a)利用结果变量的正态累积分布函数曲线，(b)求当缓解的概率等于 0.5 时所估计的变动率，(c)求在标识指数的第一个和第三个四分位点(14 和 28)处所估计的缓解概率之差。

表 6.18 习题 6.14 的数据

Parameter	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	-2.317 8	0.779 5	-4.011 4	-0.908 4	8.84	0.002 9
LI	0.087 8	0.032 8	0.027 5	0.157 5	7.19	0.007 3

译者注——LI：Labeling Index 标识指数。

6.15 利用 probit 模型来描述表 4.3 中壳宽和颜色对马蹄蟹拥有同伴的概率的效应，加以解释。

6.16 参见表 6.14。使用双对数连结而不是补余双对数连结拟合模型，检验模型的拟合情况，为什么它的拟合很差？

6.17 在表 3.9 的数据中，使用赋值 (0,15,30) 拟合线性 Logit 模型，对 $H_0:\beta=0$ 进行精确检验，并利用条件似然函数求关于 β 的点估计和区间估计，加以解释。

6.18 参见表 6.16。对第 6.7.8 节所讨论的模型进行条件 logistic 回归。

a. 以效应为正为备择假设，求 C 的效应为零的假设检验的精确 P 值。构建关于 CD 的条件发生比之比的 95% 的置信区间。

b. 在 (A,L) 的每个取值组合处，构建关于 C 和 D 的分表。注意，其中三个分表在 C = 1 时没有观测值。对于唯一一个在 C 的两个取值处都有数据的分表，求关于发生比之比的 95% 的精确置信区间以及单边精确检验的 P 值。将这些结果与使用整个数据所得到的结果进行比较。讨论那些只包括一个正的行总计或列总计的表对统计推断的贡献。

c. 对 logistic 回归模型进行普通最大似然拟合。考察有关 C 的效应估计的稳定性，在对数据增加一个观测值——即 (C,A,L) = (1,1,1) 的没有出现痢疾的案例后，求模型中 C 的效应的估计值及其标准误的变化。

6.19 表 6.19 取自一项关于非转移性骨肉瘤的研究(A. M. Goorin, *J. Clin Oncol.* 5: 1178-1184, 1987, 以及 *LogXact* 用户手册)。结果变量为研究对象是否出现了三年的无病期。

a. 显示在模型中单独包括每个预测变量时，它们的效应都是显著的。

b. 尝试拟合一个包括所有三个预测变量的主效应的 logistic 模型，说明为什么关于淋巴球渗透的效应的最大似然估计是无穷大。

表 6.19 习题 6.19 的数据

淋巴球渗透	性别	成骨细胞病理	未发病	
			是	否
高	女	否	3	0
		是	2	0
	男	否	4	0
		是	1	0
低	女	否	5	0
		是	3	2
	男	否	5	4
		是	6	11

来源：*LogXact 4 for Windows*（Cambridge，MA：CYTEL Software，1999）。

c. 利用条件 logistic 回归：(i)在控制其他变量后，进行关于淋巴球渗透的效应的精确检验；(ii)给出该效应的 95% 的置信区间。对结果加以解释。

6.20 使用本章所介绍的方法为表 5.5 选取一个模型。

6.21 在对大型金融数据的分析中，logistic 回归的应用越来越多，比如在信用评级中通过模型分析预测变量对于消费者是否值得信赖的影响。在 *www. stat. uni-muenchen. de* 网站的数据库中包含这样一个数据，它共有 1 000 个观测值和 20 个协变量。以现有账户、信用时长、信用支付史、使用目的、性别和婚姻状况为预测变量，构建一个关于信用可信度的模型。

理论与方法

6.22 在 s 个依次为 $M_1,\cdots,M_s$ 的嵌套模型中，M_s 代表最复杂的模型。令 v 表示模型 M_1 和 M_s 的残差自由度之差。

a. 说明为什么对于 $j<k$，存在 $G^2(M_j \mid M_k) \leqslant G^2(M_j \mid M_s)$。

b. 假定模型 M_j 成立，所以当 $k>j$ 时，模型 M_k 也成立。对于所有的 $k>j$，随着 $n\to\infty$，$P[G^2(M_j \mid M_k) > \chi_v^2(\alpha)] \leqslant \alpha$。说明原因。

c. Gabriel(1966)提出了一种同时对多个模型进行检验的程序，其中对于每两个模型，G^2 值之差的临界值是 $\chi_v^2(\alpha)$。最后被接受的模型必须比在两两对比中被拒绝的所有模型更复杂。由于(b)部分对于所有 $j<k$ 都成立，说明 Gabriel 程序出现第一类错误的概率不大于 α。

6.23 证明关于 $I\times 2$ 表格的线性 Logit 模型的皮尔逊残差满足 $X^2=\sum_{i=1}^{I} e_i^2$。指出这对于使用任意连结的二项分布广义线性模型都成立。

6.24 参考关于 $2\times 2\times K$ 列联表 $\{n_{ijk}\}$ 的式 6.4 Logit 模型。

a. 利用虚拟变量，写出模型的对数似然函数，指出各个参数的充分统计量，说明在控制 Z 后，如何构建关于 X 的效应的精确置信区间。

b. 利用有关指数分布族检验的基本结果，说明为什么关于 XY 的条件独立性的一致最强无偏检验(uniformly most powerful unbiased tests)需要以 $\sum_k n_{11k}$ 为基础(Birch，1964b；Lehmann，1986，Sec. 4.8)。

6.25 假定在一个 $2\times2\times2$ 表中,当 $Z=1$ 时分行的 $\{\pi_{ijk}\}$ 为 $(0.15,0.10/0.10,0.15)$,当 $Z=2$ 时分行的 $\{\pi_{ijk}\}$ 为 $(0.10,0.15/0.15,0.10)$。对于在以 Y 为结果变量的 Logit 模型中,检验 XY 的条件独立性,说明为什么比较模型 $X+Z$ 和模型 Z 的似然比检验不具有一致性,但是对 XY 的条件独立性模型拟合的似然比检验具有一致性。

6.26 参考第 6.4.1 节。当 Y 服从 $N(\mu_i,\sigma^2)$ 时,基于 X 的 I 个类别上的独立样本,比较 $(\mu_1,\cdots,\mu_I)$。当大致满足 $\mu_i=\alpha+\beta x_i$ 时,说明为什么关于 $H_0:\beta=0$ 的 t 检验或 F 检验比单维方差分析(one-way ANOVA)的 F 检验效能更高。指出在 $\{\mu_i\}$ 满足什么条件的情况下,方差分析的检验更具效能。

6.27 在多项分布中,令 $\gamma=\sum_i b_i\pi_i$,并假定 $\pi_i=f_i(\theta)>0$, $i=1,\cdots,I$。对于样本比例 $\{p_i\}$,令 $S=\sum_i b_ip_i$。令 $T=\sum_i b_i\hat{\pi}_i$,其中关于 θ 的最大似然估计值 $\hat{\theta}$ 满足 $\hat{\pi}_i=f_i(\hat{\theta})$。

a. 证明 $\mathrm{var}(S)=[\sum_i b_i^2\pi_i-(\sum_i b_i\pi_i)^2]/n$。

b. 利用 δ 方法,证明 $\mathrm{var}(T)\approx[\mathrm{var}(\hat{\theta})][\sum_i b_if_i'(\theta)]^2$。

c. 通过计算 $L(\theta)=\sum_i n_i\log[f_i(\theta)]$ 的信息矩阵,证明 $\mathrm{var}(\hat{\theta})$ 近似等于 $[n\sum_i(f_i'(\theta))^2/f_i(\theta)]^{-1}$。

d. 证明 $\mathrm{var}[\sqrt{n}(T-\gamma)]\leqslant\mathrm{var}[\sqrt{n}(S-\gamma)]$ 渐近成立(*提示*:证明 $\mathrm{var}(T)/\mathrm{var}(S)$ 是两个随机变量的相关系数的平方,其中第一个变量为 b_i 以及第二个变量为 $f_i'(\theta)/f_i(\theta)$ 的概率等于 π_i。)

6.28 利用*临界值模型*(*threshold model*)也可以推导出 probit 模型。在临界值模型中,存在一个未观测到的连续结果变量 Y^*,使得当 $y_i^*\leqslant\tau$ 时观测到 $y_i=0$,当 $y_i^*>\tau$ 时观测到 $y_i=1$。假定 $y_i^*=\mu_i+\varepsilon_i$,其中 $\mu_i=\alpha+\beta x_i$,并且 $\{\varepsilon_i\}$ 服从独立的 $N(0,\sigma^2)$ 分布。为了确保模型的可识别性(identifiability),可以设定 $\sigma=1$、临界值 $\tau=0$,证明 *probit* 模型成立,并说明为什么 β 代表 x 每增加 1 单位所预期的 Y^* 的标准差变动的单位。

6.29 考虑对两个选项的选择,比如两个产品品牌。令 U_0 表示结果 $y=0$ 的*效用*(*utility*),U_1 表示结果 $y=1$ 的效用。对于 $y=0$ 和 1,假定 $U_y=\alpha_y+\beta_yx+\varepsilon_y$,通过某种转换使得 ε_y 服从标准化分布。当 $U_1>U_0$ 时,我们选择 $y=1$。

a. 如果 ε_0 和 ε_1 是独立的 $N(0,1)$ 随机变量,证明 $P(Y=1)$ 满足 probit 模型。

b. 如果 ε_y 是独立的极值分布随机变量,其累积分布函数为 $F(\varepsilon)=\exp[-\exp(-\varepsilon)]$,证明 $P(Y=1)$ 满足 logistic 回归模型(Maddala, 1983, p. 60; McFadden, 1974)。

6.30 考虑使用补余双对数连结的模型(式 6.12)。

a. 求在 $\pi(x)=\frac{1}{2}$ 时的 x 值。

b. 证明 $\pi(x)$ 的最大变动率发生在 $x=-\alpha/\beta$ 处。在这点上,$\pi(x)$ 等于多少?给出使用双对数连结模型的相应结果,并将其与 Logit 模型和 probit 模型进行比较。

6.31 假定式 6.13 的双对数模型成立,说明如何解释 β。

6.32 令 y_i 表示 n 个独立的二分随机变量,$i=1,\cdots,n$。

a. 推导 probit 模型 $\Phi^{-1}[\pi(\mathbf{x}_i)] = \sum_j \beta_j x_{ij}$ 的对数似然函数。

b. 证明 logistic 和 probit 回归模型的似然方程是

$$\sum_i (y_i - \hat{\pi}_i) z_i x_{ij} = 0, \quad j = 0, \cdots, p,$$

其中对于 logistic 模型，$z_i = 1$；对于 probit 模型，$z_i = \phi(\sum_j \hat{\beta}_j x_{ij}) / \hat{\pi}_i (1 - \hat{\pi}_i)$（当连结函数是非典型连结时，无法通过充分统计量对数据进行简化）。

6.33 有时，样本比例是连续的而不是如同（成功数）/（试验数）的二项分布形式。每个观测值都是 0 到 1 之间的任意实数，比如牙齿表面被牙斑覆盖的比例。对于独立的结果变量 $\{y_i\}$，Aitchison 和 Shen（1980）以及 Bartlett（1937）使用了模型 $\text{Logit}(Y_i) \sim N(\beta_i, \sigma^2)$。这时，$Y_i$ 本身服从 *logistic-正态分布*（*logistic-normal distribution*）。

a. 将一个 $N(\beta, \sigma^2)$ 变量表示为 $\beta + \sigma Z$，其中 Z 是标准正态分布变量，证明 $Y_i = \exp(\beta_i + \sigma Z) / [1 + \exp(\beta_i + \sigma Z)]$。

b. 证明在 σ 很小的情况下，

$$Y_i = \frac{e^{\beta_i}}{1 + e^{\beta_i}} + \frac{e^{\beta_i}}{1 + e^{\beta_i}} \frac{1}{1 + e^{\beta_i}} \sigma Z + \frac{e^{\beta_i}(1 - e^{\beta_i})}{2(1 + e^{\beta_i})^3} \sigma^2 Z^2 + \cdots。$$

c. 令 $\mu_i = e^{\beta_i} / (1 + e^{\beta_i})$，当 σ 接近于 0 时证明

$$E(Y_i) \approx \mu_i, \quad \text{var}(Y_i) \approx [\mu_i (1 - \mu_i)]^2 \sigma^2。$$

d. 对于独立的连续比例 $\{y_i\}$，令 $\mu_i = E(Y_i)$。就广义线性模型而言，使用一个关于 μ_i 的累积分布函数的反函数作为连结是合理的，但是对如何选取关于 Y_i 的分布并不是很明确。对 logistic-正态分布的矩量的近似推出了一种方差函数为 $v(\mu_i) = \phi[\mu_i(1 - \mu_i)]^2$ 的类似然（quasi-likelihood）方法（Wedderburn，1974），其中 ϕ 为未知参数。说明为什么这样做的结果与对样本 Logit 拟合一个假定方差恒定的正态回归模型相似（类似然法具有不需要对 0 或 1 的观测值进行调整的优点，对于这些值，样本 Logit 不存在）。

e. Wedderburn（1974）举了一个例子，结果变量为一片叶子上出现某种斑点的比例。想象将叶子分割成大量面积相等的小区域，并观察每个小区域是否主要被斑点覆盖，作为对二项分布的一种近似。说明为什么这意味着 $v(\mu_i) = \phi \mu_i (1 - \mu_i)$。这样做违背了二项分布的哪条假定因而可能存在问题？（β 分布的参数族具有这种形式的方差函数（参见第 13.3.1 节）。Barndorff-Nielsen 和 Jørgensen（1991）提出了一个具有 $v(\mu_i) = \phi[\mu_i(1 - \mu_i)]^3$ 的分布；另见：Cox（1996））。

6.34 对于独立的二项分布样本，构建对数似然函数，并指出对式 6.4 模型中的 β 进行精确推断需要消除掉的参数的充分统计量。

6.35 令 $\hat{\boldsymbol{\pi}}^{(-)} = (\hat{\pi}^{(-1)}, \ldots, \hat{\pi}^{(-n)})$，其中 $\hat{\pi}^{(-i)}$ 表示在拟合了一个不包括二分观测值 i 的模型后对 $E(Y_i)$ 的估计。交叉验证（cross-validation）表明，当相关系数 $\text{corr}(\hat{\boldsymbol{\pi}}^{(-)}, \mathbf{y})$ 较大时，模型具有较好的预测力。对于所有 i，考虑模型 $\text{Logit}(\pi_i) = \alpha$，证明 $\hat{\pi}_i = \bar{y}$，进而存在 $\hat{\pi}^{(-i)} = [n/(n-1)][\bar{y} - (1/n) y_i]$，而且无论模型的拟合结果如何，都有 $corr(\hat{\boldsymbol{\pi}}^{(-)}, \mathbf{y}) = -1$。因此，关于二分数据的交叉验证可能会得出误导性结论（*Zheng and Agresti*，2000）。

7 关于多项结果变量的Logit 模型

在第 5 章和第 6 章,我们讨论了关于二分结果变量的二项分布广义线性模型。对于多类别的结果变量,需要应用多项分布广义线性模型。在本章中,我们将 logistic 回归扩展到多项(定类和定序)结果变量的情况。

第 7.1 节介绍一个关于定类结果变量的模型,它对结果变量的每一对类别分别拟合二分的 Logit 模型。第 7.2 节介绍关于定序结果变量的模型,它应用的是结果变量的累积分布概率的 Logit。在第 7.3 节中,我们给出这些累积概率的其他连结函数。第 7.4 节讨论其他可选择的定序结果变量模型。

在第 7.5 节中,我们讨论利用模型或一般化的 Cochran-Mantel-Haenszel 统计量对多项结果变量进行条件独立性检验。在本章最后一节,我们介绍关于*离散选择*(*discrete-choice*)的多项 Logit 模型,即研究对象从多个选项中做出选择,而且预测变量的取值可以随着选项的不同而不同。

7.1 定类结果变量:基线类别 Logit 模型

令 Y 表示一个具有 J 个类别的分类结果变量。关于定类结果变量的多类别(也称为*多项*(*polytomous*))Logit 模型同时描述所有$\binom{J}{2}$个类别配对的对数发生比。按照某种方式从中选出 $J-1$ 个对数发生比,剩下的都是冗余的。

7.1.1 基线类别 Logit

令 $\pi_j(\mathbf{x}) = P(Y=j \mid \mathbf{x})$ 表示在解释变量的给定取值 $\mathbf{x}$ 处结果变量取值为 j 的概率,存在 $\sum_j \pi_j(\mathbf{x}) = 1$ 。对于位于 $\mathbf{x}$ 处的观测值,我们将 Y 落在这 J 个类别的计数视为一个多项分布,其概率等于 $\{\pi_1(\mathbf{x}), \cdots, \pi_J(\mathbf{x})\}$ 。

通常将结果变量的最后一个或比例最大的一个类别作为基线类别(baseline category),Logit 模型将结果变量的每个类别与基线类别进行配对。模型

$$\log \frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} = \alpha_j + \boldsymbol{\beta}_j'\mathbf{x}, \quad j = 1, \cdots, J-1, \tag{7.1}$$

同时描述了 $\mathbf{x}$ 对这些 $J-1$ 个 Logit 的效应。这些效应随着结果变量中与基线类别配对的类别的变化而变化。这 $J-1$ 个方程式决定了结果变量任意两个类别之间的 Logit 参数,因为

$$\log \frac{\pi_a(\mathbf{x})}{\pi_b(\mathbf{x})} = \log \frac{\pi_a(\mathbf{x})}{\pi_J(\mathbf{x})} - \log \frac{\pi_b(\mathbf{x})}{\pi_J(\mathbf{x})}。$$

当预测变量为分类变量且不存在稀疏数据问题时,可以用 X^2 和 G^2 拟合优度统计量来检查模型的拟合。当存在连续的解释变量或出现稀疏数据的情况时,这些统计量仍然可以用来对参数个数相差不多的嵌套模型进行有效的比较(Haberman,1974a,pp. 372-373;1977a)。

7.1.2　例子:短吻鳄的觅食

表 7.1 取自一项关于短吻鳄主要觅食选择的影响因素的研究。该研究的对象是在佛罗里达的四个湖泊里捕获的 219 只短吻鳄。定类结果变量为在短吻鳄的胃中发现的主要食物类型。这些食物共包括 5 个类别:鱼类、无脊椎动物、爬行动物、鸟类,以及其他。无脊椎动物包括福寿螺、水生昆虫以及小龙虾。爬行动物主要是指乌龟,但在其中一只短吻鳄的胃里还发现了前一年放生到湖里的 23 只幼鳄的标签!"其他"类主要包括两栖动物、哺乳动物、植物、石头或其他的残骸、没有食物或食物种类没有明显特征。表 7.1 还按照 L = 被捕获的湖泊(汉考克湖,奥克拉沃霍湖,特拉福德湖,乔治湖)、G = 性别(雌性,雄性)以及 S = 尺寸(≤2.3 米, > 2.3 米)对短吻鳄进行了划分。

表 7.1　短吻鳄的主要觅食类型

湖　泊	性　别	尺　寸/m	主要食物选择				
			鱼　类	无脊椎动物	爬行动物	鸟　类	其　他
汉考克	雄性	≤2.3	7	1	0	0	5
		>2.3	4	0	0	1	2
	雌性	≤2.3	16	3	2	2	3
		>2.3	3	0	1	2	3
奥克拉沃霍	雄性	≤2.3	2	2	0	0	1
		>2.3	13	7	6	0	0
	雌性	≤2.3	3	9	1	0	2
		>2.3	0	1	0	1	0
特拉福德	雄性	≤2.3	3	7	1	0	1
		>2.3	8	6	6	3	5
	雌性	≤2.3	2	4	1	1	4
		>2.3	0	1	0	0	0
乔治	雄性	≤2.3	13	10	0	2	2
		>2.3	9	0	0	1	2
	雌性	≤2.3	3	9	1	0	1
		>2.3	8	1	0	0	1

来源:数据取自 M. F. Delaney 和 C. T. Moore 的未发表文稿,由 Clint Moore 友情提供。

利用基线类别 Logit 模型可以考察 L, G 和 S 对短吻鳄的主要觅食类型的效应。表 7.2给出了几个模型的拟合统计量。我们通过模型所包括的预测变量的符号来表示每个模型:例如,模型 $(L+S)$ 包括湖泊和尺寸的可加效应,而模型( )不包括任何预测变量。这里存在数据稀疏的问题,219 个观测值散落在 80 个单元格中。因此,用 G^2 进行模型比较比用它来检验拟合更可靠。统计量 $G^2[(\ )\mid(G)] = 2.1$, $G^2[(L+S)\mid(G+L+S)] = 2.2$,二者的自由度均为 df = 4。结果表明,应该将性别进行合并来简化表格(文中未显示的其他分析表明,加入包括 G 的交互项也不能显著地改善模型的拟合结果)。简化后,表

格的 G^2 和 X^2 值显示，L 和 S 都具有显著效应。表 7.3 给出了对简化表格拟合模型（$L+S$）的结果。对于该表的 40 个单元格，其中只有两个用于比较观察值和拟合值的标准化皮尔逊残差的绝对值超过了 2，所有残差都不超过 3。拟合结果看上去是充分的。

表 7.2　关于表 7.1 数据的基线类别 Logit 模型的拟合优度

模　型[a]	G^2	X^2	df
(　)	116.8	106.5	60
(G)	114.7	101.2	56
(S)	101.6	86.9	56
(L)	73.6	79.6	48
($L+S$)	52.5	58.0	44
($G+L+S$)	50.3	52.6	40
合并 G			
(　)	81.4	73.1	28
(S)	66.2	54.3	24
(L)	38.2	32.7	16
($L+S$)	17.1	15.0	12

a G，性别；S，尺寸；L，被捕获的湖泊。有关细节，参见正文。

表 7.3　短吻鳄觅食研究的观测值和拟合值

湖　泊	短吻鳄的尺寸/m	主要食物选择				
		鱼　类	无脊椎动物	爬行动物	鸟　类	其　他
汉考克	≤2.3	23	4	2	2	8
		(20.9)	(3.6)	(1.9)	(2.7)	(9.9)
	>2.3	7	0	1	3	5
		(9.1)	(0.4)	(1.1)	(2.3)	(3.1)
奥克拉沃霍	≤2.3	5	11	1	0	3
		(5.2)	(12.0)	(1.5)	(0.2)	(1.1)
	>2.3	13	8	6	1	0
		(12.8)	(7.0)	(5.5)	(0.8)	(1.9)
特拉福德	≤2.3	5	11	2	1	5
		(4.4)	(12.4)	(2.1)	(0.9)	(4.2)
	>2.3	89	7	6	3	5
		(8.6)	(5.6)	(5.9)	(3.1)	(5.8)
乔治	≤2.3	16	19	1	2	3
		(18.5)	(16.9)	(0.5)	(1.2)	(3.8)
	>2.3	17	1	0	1	3
		(14.5)	(3.1)	(0.5)	(1.8)	(2.2)

鱼类是最普遍的食物选择。现在，我们估计湖泊和尺寸对短吻鳄选择其他种类的食物而不是鱼类的发生比的效应。以鱼类作为基线类别，表 7.4 给出了相应效应参数的最大似然估计。这些结果来自于用虚拟变量分别表示前三个湖泊以及尺寸的模型。表中使用字母下标来表示食物选择的类别。例如，对于选择无脊椎动物而不是鱼类的对数发生比的预测方程为

$$\log(\hat{\pi}_I/\hat{\pi}_F) = -1.55 + 1.46s - 1.66z_H + 0.94z_O + 1.12z_T,$$

其中 $s=1$ 表示尺寸 $\leqslant 2.3$ 米，否则为 0；z_H 是关于汉考克湖的虚拟变量（即 $z_H=1$ 表示该湖的短吻鳄，否则为 0），z_O 和 z_T 分别是关于奥克拉沃霍湖和特拉福德湖的虚拟变量。

短吻鳄的尺寸对于觅食选择具有很明显的效应。对于给定的一个湖泊，所估计的小短吻鳄的主要食物为无脊椎动物而不是鱼类的发生比是大短吻鳄的 $\exp(1.46) = 4.3$ 倍；该效应的沃尔德置信区间为 $\exp[1.46 \pm 1.96(0.396)] = (2.0, 9.3)$。湖泊的效应表明，在特拉福德湖和奥克拉沃霍湖所估计的主要食物是无脊椎动物而不是鱼类的发生比乔治湖相对较大，而汉考克湖则比乔治湖相对较小。

表 7.4 通过虚拟变量表示尺寸的第一类别以及除乔治湖外的其他湖泊，关于短吻鳄觅食种类的 Logit 模型的参数估计[a]

Logit[b]	截　距	尺寸≤2.3	湖　泊		
			汉考克	奥克拉沃霍	特拉福德
$\text{Log}(\pi_I/\pi_F)$	-1.55	1.46(0.40)	-1.66(0.61)	0.94(0.47)	1.12(0.49)
$\text{Log}(\pi_R/\pi_F)$	-3.31	-0.35(0.58)	1.24(1.19)	2.46(1.12)	2.94(1.12)
$\text{Log}(\pi_B/\pi_F)$	-2.09	-0.63(0.64)	0.70(0.78)	-0.65(1.20)	1.09(0.84)
$\text{Log}(\pi_O/\pi_F)$	-1.90	0.33(0.45)	0.83(0.56)	0.01(0.78)	1.52(0.62)

a 括号中数值为标准误

b I，无脊椎动物；R，爬行动物；B，鸟类；O，其他；F，鱼类。

表 7.4 中的方程还决定了不同觅食种类之间的其他配对。例如，对于(无脊椎动物，其他)，

$$
\begin{aligned}
\log(\hat{\pi}_I/\hat{\pi}_O) &= \log(\hat{\pi}_I/\hat{\pi}_F) - \log(\hat{\pi}_O/\hat{\pi}_F) \\
&= (-1.55 + 1.46s - 1.66z_H + 0.94z_O + 1.12z_T) - \\
&\quad (-1.90 + 0.33s + 0.83z_H + 0.01z_O + 1.52z_T) \\
&= 0.35 + 1.13s - 2.48z_H + 0.93z_O - 0.39z_T。
\end{aligned}
$$

7.1.3 估计结果变量的发生概率

将多项 Logit 模型直接表示为结果变量发生概率 $\{\pi_j(\mathbf{x})\}$ 的公式为

$$
\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}'_j\mathbf{x})}{1 + \sum_{h=1}^{J-1}\exp(\alpha_h + \boldsymbol{\beta}'_h\mathbf{x})}, \tag{7.2}
$$

其中 $\alpha_J = 0$ 且 $\boldsymbol{\beta}_J = 0$。这可以由式 7.1 推出，对于 $j = J$，将 α_J 和 $\boldsymbol{\beta}_J$ 设定为零，式7.1仍然成立(同时，为了保证模型的可识别性，要求基线类别的参数等于零；参见习题7.26)。式 7.2 的分母对于每个 j 都一样。所有 j 的分子加起来等于分母，所以 $\sum_j \pi_j(\mathbf{x}) = 1$。当 $J = 2$ 时，式 7.2 简化为二分 logistic 回归对应的式 5.1 的形式。

利用表 7.4，在汉考克湖的一只大短吻鳄主要觅食无脊椎动物的概率估计值等于

$$
\hat{\pi}_I = \frac{e^{-1.55-1.66}}{1 + e^{-1.55-1.66} + e^{-3.31+1.24} + e^{-2.09+0.70} + e^{-1.90+0.83}} = 0.023。
$$

主要觅食爬虫动物、鸟类、其他和鱼类的概率估计值分别为 0.072，0.141，0.194 和 0.570。

这个例子所使用的预测变量是分类变量。多项 Logit 模型也可以包括连续的预测变量。在这项研究中，生物学家利用关于尺寸的虚拟变量来区分成年的和幼年的短吻鳄。但是，该研究其实测量了短吻鳄的实际长度，而它是一个连续变量。在预测变量为连续变量的情况下，将估计的概率进行绘图能提供许多信息。具体来说，对于某个湖中的短吻鳄，图 7.1 将主要食物类型为鱼类、无脊椎动物或其他(将另外三类加以合并)的估计概率分别表示为长度的函数。在结果变量具有多个类别的情况下，某个给定类别的概率

不一定会持续上升或下降(习题7.27)。

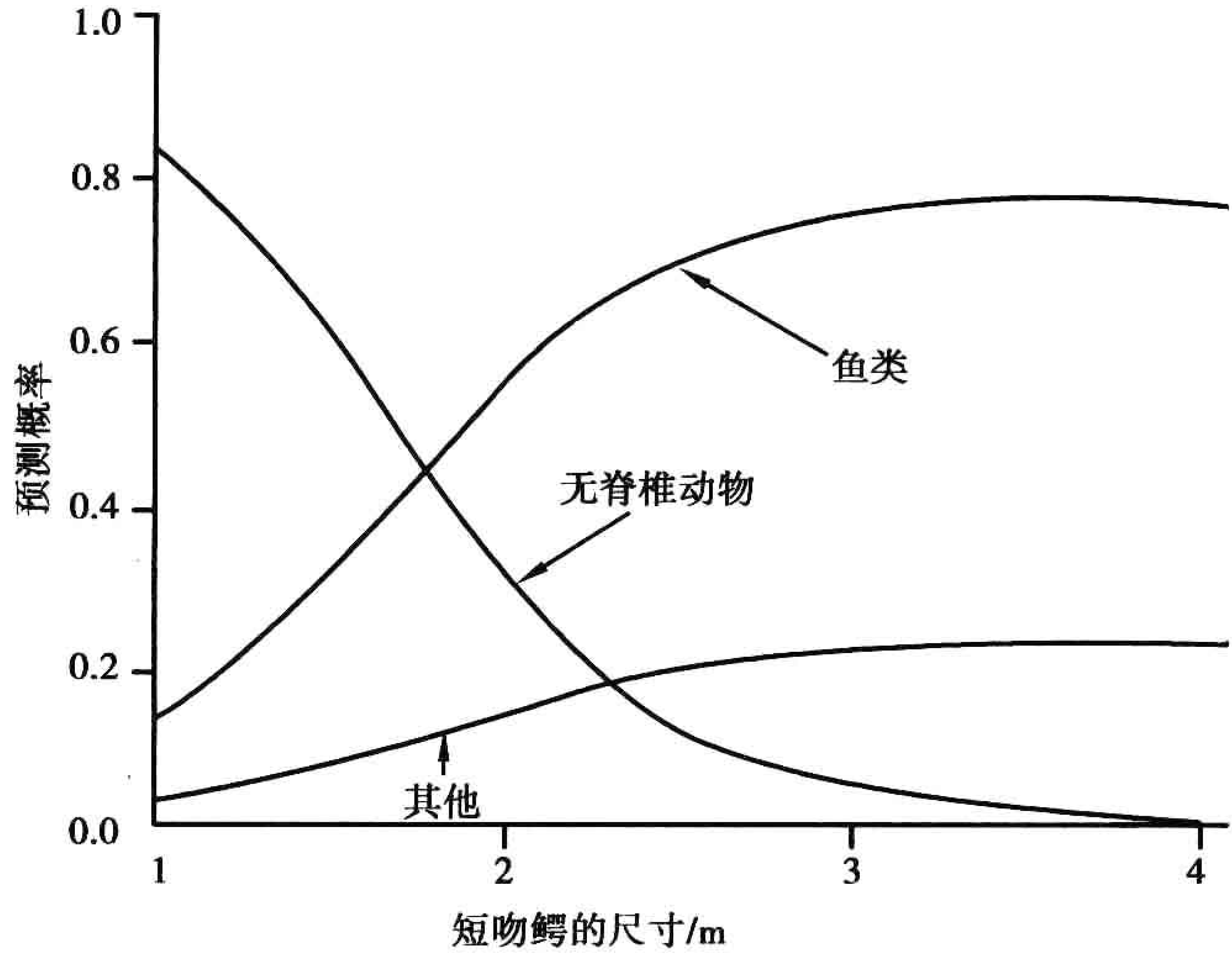


图7.1 主要觅食类型的估计概率

7.1.4 基线类别 Logit 模型的拟合*

关于多项 Logit 模型的最大似然拟合在 $\{\pi_j(\mathbf{x})\}$ 同时满足模型设定的 $J-1$ 个方程的限定条件下最大化似然函数。对于 $i=1,\cdots,n$，令 $\mathbf{y}_i=(y_{i1},\cdots,y_{iJ})$ 表示关于第 i 个对象的多项分布试验,其中当试验结果落在类别 j 时 $y_{ij}=1$，否则 $y_{ij}=0$。因此，$\sum_j y_{ij}=1$。令 $x_i=(x_{i1},\cdots,x_{ip})'$ 表示第 i 个对象对应的解释变量取值，$\boldsymbol{\beta}_j=(\beta_{j1},\cdots,\beta_{jp})'$ 代表关于第 j 个 Logit 的参数。

由于 $\pi_J=1-(\pi_1+\cdots+\pi_{J-1})$ 且 $y_{iJ}=1-(y_{i1}+\cdots+y_{i,J-1})$，在对数似然函数中对象 i 的贡献为

$$\log\left[\prod_{j=1}^{J}\pi_j(\mathbf{x}_i)^{y_{ij}}\right]=\sum_{j=1}^{J-1}y_{ij}\log\pi_j(\mathbf{x}_i)+\left(1-\sum_{j=1}^{J-1}y_{ij}\right)\log\left[1-\sum_{j=1}^{J-1}\pi_j(\mathbf{x}_i)\right]$$

$$=\sum_{j=1}^{J-1}y_{ij}\log\frac{\pi_j(\mathbf{x}_i)}{1-\sum_{j=1}^{J-1}\pi_j(\mathbf{x}_i)}+\log\left[1-\sum_{j=1}^{J-1}\pi_j(\mathbf{x}_i)\right]。$$

因而,基线类别 Logit 是多项分布的自然参数。

现在,假定有 n 个独立的观测值。在上面的最后一个表达式中,将第一项中的 Logit 代入 $\alpha_j+\boldsymbol{\beta}'_j\mathbf{x}_i$ 并在第二项中代入 $\pi_J(\mathbf{x}_i)=1/\left[1+\sum_{j=1}^{J-1}\exp(\alpha_j+\boldsymbol{\beta}'_j\mathbf{x}_i)\right]$，对数似然函数可表示为

$$\log\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\pi_j(\mathbf{x}_i)^{y_{ij}}\right]$$

$$=\sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1}y_{ij}(\alpha_j+\boldsymbol{\beta}_j'\mathbf{x}_i)-\log\left[1+\sum_{j=1}^{J-1}\exp(\alpha_j+\boldsymbol{\beta}_j'\mathbf{x}_i)\right]\right\}$$

$$=\sum_{j=1}^{J-1}\left[\alpha_j\left(\sum_{i=1}^{n}y_{ij}\right)+\sum_{k=1}^{p}\beta_{jk}\left(\sum_{i=1}^{n}x_{ik}y_{ij}\right)\right]-$$

$$\sum_{i=1}^{n}\log\left[1+\sum_{j=1}^{J-1}\exp(\alpha_j+\boldsymbol{\beta}_j'\mathbf{x}_i)\right]。$$

关于 β_{jk} 的充分统计量为 $\sum_i x_{ik}y_{ij}$，$j=1,\cdots,J-1$，$k=1,\cdots,p$。在 $x_{i0}=1$ 的情况下，α_j

的充分统计量为 $\sum_i y_{ij} = \sum_i x_{i0} y_{ij}$，这等于结果落在类别 j 的总数。

似然方程设这些充分统计量等于它们的期望值。对数似然函数是一个凹函数，可以通过 Newton-Raphson 法求得参数的最大似然估计。这些估计值服从大样本正态分布，它们的渐近标准误等于逆信息矩阵中对角线元素的平方根。

大多数统计软件都可以拟合多项 Logit 模型，但也有一些软件只能拟合二分 logistic 回归模型。另一种可考虑的拟合方法是，对结果变量类别之间的 $J-1$ 个配对分别拟合二分 Logit 模型：只使用结果变量落在第 1 类或第 J 类的观测值，单独对 $j=1$ 拟合式 7.1 模型，从而获得关于 α_1 和 $\boldsymbol{\beta}_1$ 的估计；使用第 2 类和第 J 类的观测值来获得关于 α_2 和 $\boldsymbol{\beta}_2$ 的估计；以此类推，分别获取 $J-1$ 个 Logit 模型的拟合结果。仅仅使用结果变量的两个类别所拟合的 Logit 模型与以这种类别划分为条件所拟合的普通 Logit 模型是相同的。例如，第 j 个基线类别 Logit 是一个关于条件概率的 Logit：

$$\log \frac{\pi_j(\mathbf{x})/(\pi_j(\mathbf{x}) + \pi_J(\mathbf{x}))}{\pi_J(\mathbf{x})/(\pi_j(\mathbf{x}) + \pi_J(\mathbf{x}))} = \log \frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})}。$$

分别拟合的估计结果不同于对 $J-1$ 个 Logit 同时进行最大似然拟合。前者效率更低，一般会具有较大的标准误。不过，Begg 和 Gray(1984)表明，当以结果变量所占比例最大的类别作为基线时，效率的损失有限。以单独分析只包括无脊椎动物和鱼类的数据为例，我们来展示一下这种方法。拟合结果为 $\log(\hat{\pi}_I/\hat{\pi}_F) = -1.69 + 1.66s - 1.78z_H + 1.05z_O + 1.22z_T$，效应的标准误分别为 (0.43, 0.62, 0.49, 0.52)。这一模型对效应的拟合结果与同时拟合所有 5 个类别的结果相似——参见表 7.4 的第一行。它所估计的标准误只是略大一些，因为在 219 个观测值中有 155 个的觅食类型为鱼类或无脊椎动物。

7.1.5 作为多元广义线性模型的多类别 Logit 模型*

对于服从自然指数分布族的一个结果变量，广义线性模型具有 $g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$ 的形式。其中，连结函数为 g，结果变量的期望 $\mu_i = E(Y_i)$，第 i 个观测值对应的 p 个解释变量的取值向量为 $\mathbf{x}_i$，相应参数向量 $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$。在多元指数分布族中，如多项分布，该模型扩展为一个多元广义线性模型(习题 7.24)。

令 $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots)'$ 表示第 i 个对象的结果变量所组成的向量，存在 $\boldsymbol{\mu}_i = E(\mathbf{Y}_i)$。令 $\mathbf{g}$ 表示连结函数的向量。多元广义线性模型具有以下形式：

$$\mathbf{g}(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta}, \tag{7.3}$$

其中第 i 个观测值的模型矩阵 $\mathbf{X}_i$ 的第 h 行包含了关于 y_{ih} 的解释变量的取值。有关细节，参见：Fahrmeir and Tutz(2001, Chap. 3)。

基线类别 Logit 模型是一个多元广义线性模型。这里，由于 y_{iJ} 是冗余的，所以 $\mathbf{y}_i = (y_{i1}, \cdots, y_{i,J-1})'$。这时，$\boldsymbol{\mu}_i = (\pi_1(\mathbf{x}_i), \cdots, \pi_{J-1}(\mathbf{x}_i))'$，并且

$$g_j(\boldsymbol{\mu}_i) = \log\{\mu_{ij}/[1 - (\mu_{i1} + \cdots + \mu_{i,J-1})]\}。$$

第 i 个观测值的模型矩阵为

$$\mathbf{X}_i = \begin{pmatrix} 1 & \mathbf{X}_i' & & & & & \\ & & 1 & \mathbf{X}_i' & & & \\ & & & & \cdots & & \\ & & & & & 1 & \mathbf{X}_i' \end{pmatrix}。$$

该矩阵其他位置的元素为 0，并且 $\boldsymbol{\beta}' = (\alpha_1, \boldsymbol{\beta}'_1, \cdots, \alpha_{J-1}, \boldsymbol{\beta}'_{J-1})$。在分组数据的情况下，大家也可以通过每个类别的样本比例来对其进行表述。

7.2 定序结果变量:累积 Logit 模型

在第6.4.1节,我们介绍了利用定序变量的排序特征对单个参数进行推断所具有的优点。这对定序结果变量的模型分析同样适用。在模型中,具有排序特征的项,比如单调趋势,可以增进模型的简约性并提高统计效能。在本节中,我们介绍关于定序结果变量的最常见的 Logit 模型。

7.2.1 累积 Logit

利用类别间的排序特征的一种方式是计算以下累积概率的 Logit,

$$P(Y \leqslant j \mid \mathbf{x}) = \pi_1(\mathbf{x}) + \cdots + \pi_j(\mathbf{x}), \quad j = 1,\cdots,J。$$

我们将*累积 Logit*(*cumulative Logits*)定义为

$$\text{Logit}[P(Y \leqslant j \mid \mathbf{x})] = \log \frac{P(Y \leqslant j \mid \mathbf{x})}{1 - P(Y \leqslant j \mid \mathbf{x})}$$

$$= \log \frac{\pi_1(\mathbf{x}) + \cdots + \pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \cdots + \pi_J(\mathbf{x})}, j = 1,\cdots,J-1。 \quad (7.4)$$

每个累积 Logit 都使用了结果变量的所有 J 个类别。

关于 Logit $[P(Y \leqslant j)]$ 本身的模型是一个关于二分结果变量的普通 Logit 模型,其中类别1到j组成第一种结果,类别$j+1$到J组成第二种结果。更好的是,可以利用一个简约模型分析所有 $J-1$ 个累积 Logit。

7.2.2 比例发生比模型

一个同时分析所有累积 Logit 的模型可表示为

$$\text{Logit}[P(Y \leqslant j \mid \mathbf{x})] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \quad j = 1,\cdots,J-1。 \quad (7.5)$$

其中,每个累积 Logit 都具有单独的截距。$\{\alpha_j\}$ 随着j的上升而上升,因为对于给定的 $\mathbf{x}$,$P(Y \leqslant j \mid \mathbf{x})$ 随着j的上升而上升,并且 Logit 是关于这个概率的一个增函数。

这个模型中,每个 Logit 具有相同的效应 $\boldsymbol{\beta}$。对于一个连续的预测变量 x,图7.2显示了当 $J=4$ 时的累积 Logit 模型。对于给定的j,结果变量的曲线是一条关于二分结果变量的 logistic 回归曲线,两种结果分别为 $Y \leqslant j$ 和 $Y > j$。代表$j = 1,2$ 和 3 的结果变量曲线具有相同的形状。它们上升和下降的速度完全相同,只是在水平位置上有所不同。对于$j < k$,表示 $P(Y \leqslant k)$ 的曲线相当于表示 $P(Y \leqslant j)$ 的曲线在 x 的方向上平移了 $(\alpha_k - \alpha_j)/\beta$ 个单位;也即,

$$P(Y \leqslant k \mid X = x) = P(Y \leqslant j \mid X = x + (\alpha_k - \alpha_j)/\beta)。$$

图7.3显示了每个类别的概率曲线。

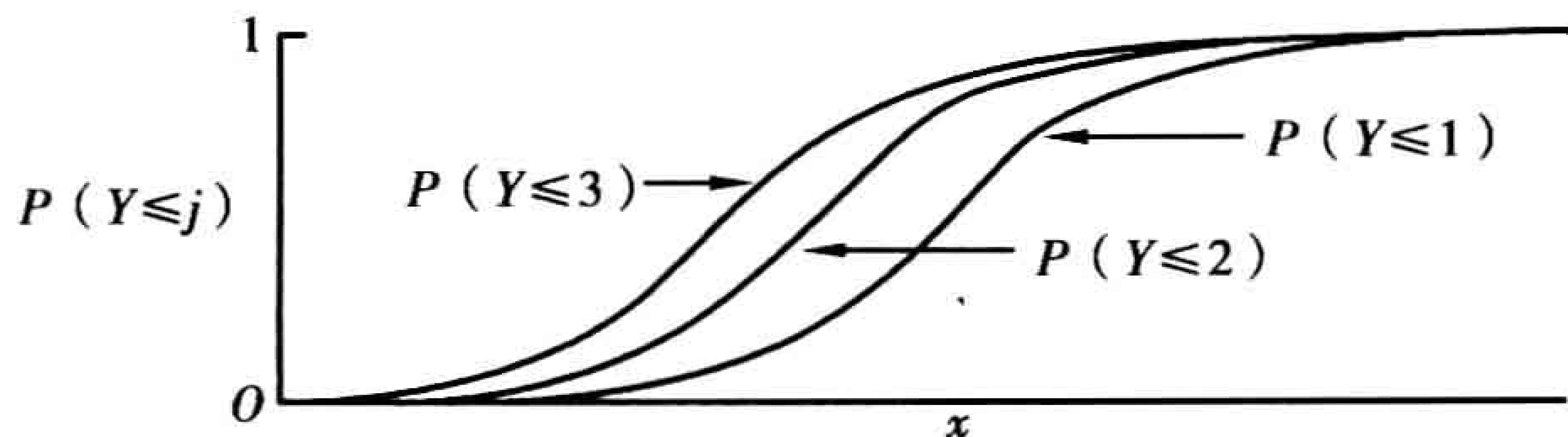


图7.2 效应与切点无关的累积 Logit 模型

式7.5累积 Logit 模型满足

$$\text{Logit}[P(Y \leqslant j \mid \mathbf{x}_1)] - \text{Logit}[P(Y \leqslant j \mid \mathbf{x}_2)]$$

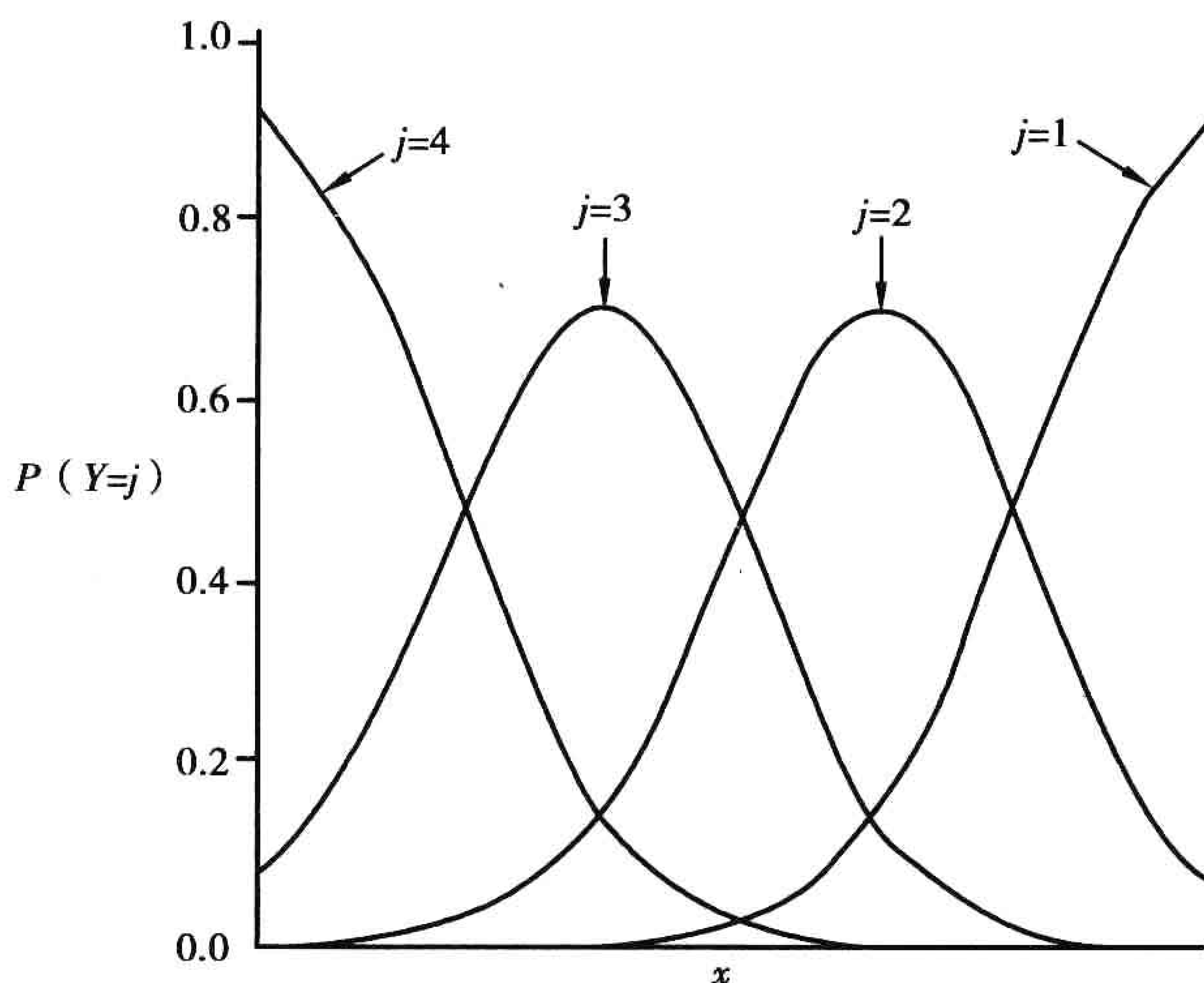


图 7.3 累积 Logit 模型中结果变量各个类别的概率

$$= \log \frac{P(Y \leqslant j \mid \mathbf{x}_1)/P(Y > j \mid \mathbf{x}_1)}{P(Y \leqslant j \mid \mathbf{x}_2)/P(Y > j \mid \mathbf{x}_2)} = \boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)。$$

累积概率之间的发生比之比被称为*累积发生比之比*（*cumulative odds ratio*）。在 $\mathbf{x} = \mathbf{x}_1$ 处结果变量的取值 $\leqslant j$ 的发生比，是在 $\mathbf{x} = \mathbf{x}_2$ 处的相应发生比的 $\exp[\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)]$ 倍。对数累积发生比之比与 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 的距离成比例。同样的等比例常数对每个 Logit 都适用。由于这个特性，McCullagh(1980)将式 7.5 称为*比例发生比模型*（*proportional odds model*）。

在只有一个预测变量的情况下，只要 $x_1 - x_2 = 1$，累积发生比之比都等于 e^{β}。图7.4演示了对于所有的 j，这个模型都隐含着相同的累积发生比之比。它显示了结果变量的 J 个类别被合并为一个二分变量（$\leqslant j$, $> j$）的情况，并给出了在这种情况下决定累积发生比之比 AD/BC 取相同值 e^{β} 的一组单元格。

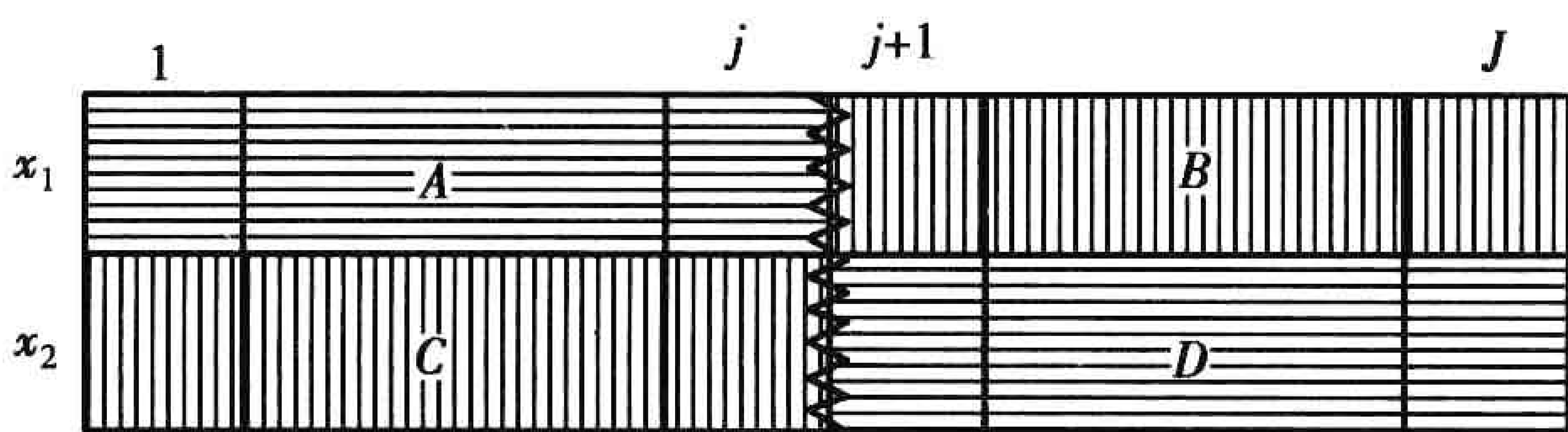


图 7.4 在比例发生比模型中，结果变量的所有切点在 $x_1 - x_2 = 1$ 时都具有相同的发生比之比 AD/BC

式 7.5 模型限定 $J-1$ 条结果变量曲线具有相同的形状，因此，它的拟合结果与对每个 j 分别拟合 Logit 模型不同。再次令 $(y_{i1}, \cdots, y_{iJ})$ 表示第 i 个对象的结果变量的二分指示值，相应的似然函数等于

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\pi_j(\mathbf{x}_i)^{y_{ij}}\right] = \prod_{i=1}^{n}\left[\prod_{j=1}^{J}(P(Y \leqslant j \mid \mathbf{x}_i) - P(Y \leqslant j-1 \mid \mathbf{x}_i))^{y_{ij}}\right]$$

$$= \prod_{i=1}^{n}\left[\prod_{j=1}^{J}\left(\frac{\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}\right)^{y_{ij}}\right], \quad (7.6)$$

它是关于$(\{\alpha_j\}, \boldsymbol{\beta})$的一个函数。McCullagh(1980)、Walker 和 Duncan(1967)利用 Fisher 计分法对它进行了最大似然估计。

7.2.3 潜变量的推理*

假定 Y 背后隐含着一个连续变量，其回归模型可推导出对于不同的 j 存在共同效应 $\boldsymbol{\beta}$

的比例发生比模型(Anderson and Philips,1981)。令 Y^* 表示这个隐含的变量。在统计学中,这种无法观测的变量被称为*潜变量*(*latent variable*)。假定它的累积分布函数为 $G(y^*-\eta)$,其中 y^* 的值围绕一个位置参数 η(比如均值)变动,η 通过 $\eta(\mathbf{x})=\boldsymbol{\beta}'\mathbf{x}$ 而取决于 $\mathbf{x}$。假定 $-\infty=\alpha_0<\alpha_1<\cdots<\alpha_J=\infty$ 为该连续刻度的*切点*(*cutpoints*),使得观察到的结果变量满足:

$$\text{如果 } \alpha_{j-1}<Y^*\leqslant\alpha_j,\quad Y=j。$$

也即,当潜变量的取值落在第 j 个区间时(图 7.5)时,Y 落在第 j 个类别。这时,

$$P(Y\leqslant j\mid\mathbf{x})=P(Y^*\leqslant\alpha_j\mid\mathbf{x})=G(\alpha_j-\boldsymbol{\beta}'\mathbf{x})。$$

关于 Y 的适当模型意味着对 $P(Y\leqslant j\mid\mathbf{x})$ 使用连结 G^{-1},即 Y^* 的累积分布函数的反函数。如果 $Y^*=\boldsymbol{\beta}'\mathbf{x}+\varepsilon$,其中 ε 的累积分布函数 G 是 logistic 函数(第 4.2.5 节),那么 G^{-1} 就是 Logit 连接,因而得到比例发生比模型。如果 ε 服从正态分布,则 G^{-1} 意味着关于累积概率的 probit 连结(第 7.3.1 节)。

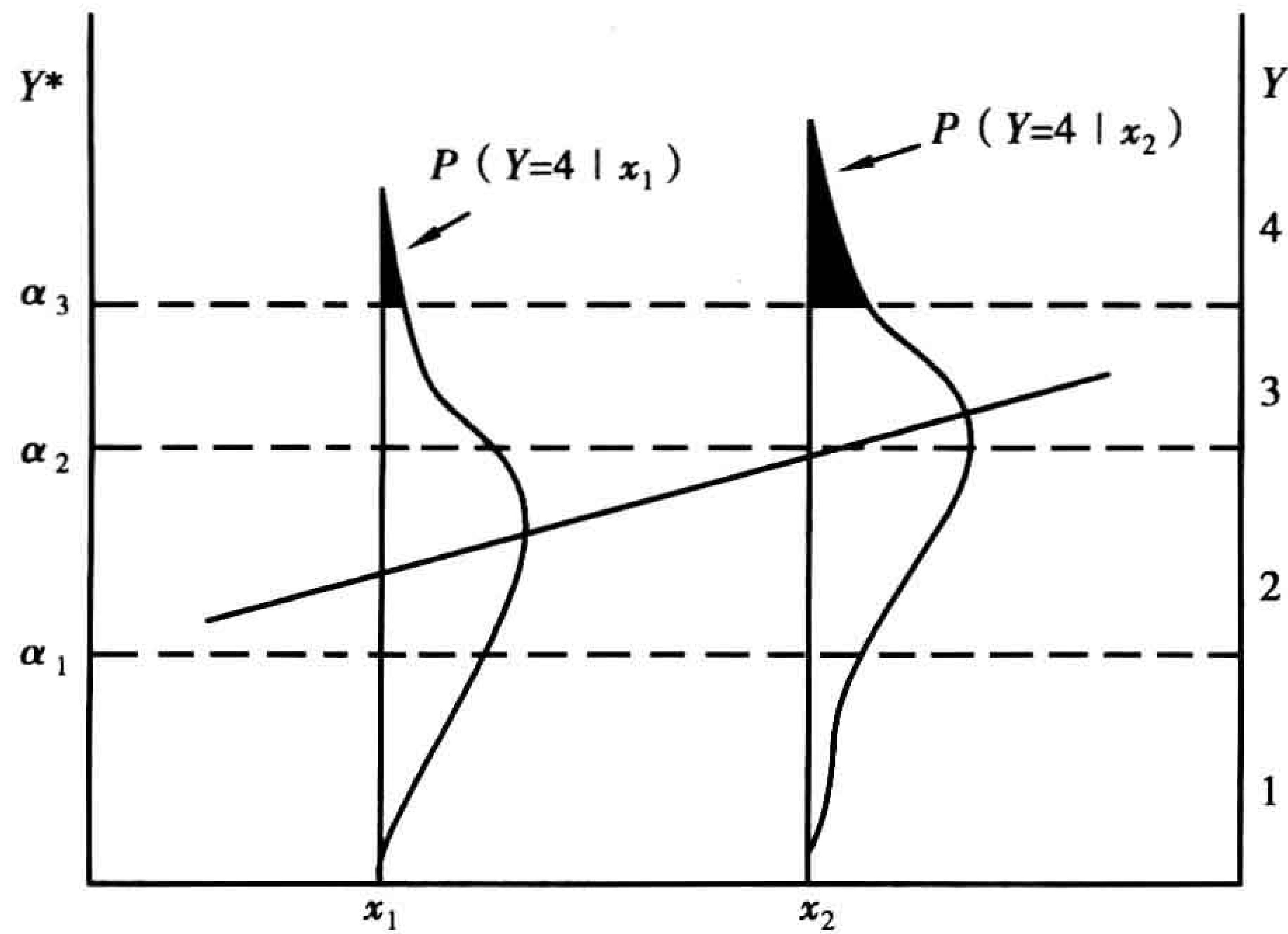


图 7.5　定序测量与潜变量所隐含的回归模型

在这个推导过程中,不管切点 $\{\alpha_j\}$ 发生在潜变量的哪一处,对 Y 的效应参数 $\boldsymbol{\beta}$ 都是相等的。效应参数不随着 Y 的分类标准的变化而变化。如果一个测量政治态度的连续变量与预测变量之间存在一个线性回归,那么同样的效应参数也适用于对政治态度的离散形式的测量,比如类别为(激进,中间,保守)或(非常激进,比较激进,中间,比较保守,非常保守)。这个特点使得我们有可能对使用不同尺度的结果变量的研究结果进行比较。

注意,对潜变量使用形式为 $G(y^*-\eta)$ 的累积分布函数所形成的线性预测项为 $\alpha_j-\boldsymbol{\beta}'\mathbf{x}$,而不是 $\alpha_j+\boldsymbol{\beta}'\mathbf{x}$。当 $\beta>0$ 时,随着 x 的上升,每个累积 Logit 会下降,所以每个累积概率也下降,并且落在 Y 的刻度低端的概率相对更小。因此,Y 在 x 取较大的值时也会比较大。按照这种参数化方式,β 的符号具有通常的意义。但是,大多数软件(如 SAS)使用的是式 7.5 的形式。

7.2.4　例子:精神损伤的研究

表 7.5 取自一项关于精神健康的研究,研究对象为佛罗里达州阿拉楚阿县(Alachua County,Florida)成年居民的一个随机样本。该表给出了精神损伤与两个解释变量的关系。其中,精神损伤是一个定序变量,它的类别为(好,轻微症状,中度症状,受损);生命事件指数 x_1 是一个关于重大生命事件(比如孩子出生、变换工作、离婚或过去 3 年内家庭

成员的死亡等)所发生的次数和严重程度的综合测量指标;社会经济地位(x_2 = SES)在这里是一个二分变量(1 = 高,0 = 低)。

表 7.5　按照社会经济地位和生命事件指数划分的精神损伤情况

研究对象	精神损伤情况	社会经济地位[a] x_2	生命事件指数 x_1	研究对象	精神损伤情况	社会经济地位[a] x_2	生命事件指数 x_1
1	好	1	1	21	轻度	1	9
2	好	1	9	22	轻度	0	3
3	好	1	4	23	轻度	1	3
4	好	1	3	24	轻度	1	1
5	好	0	2	25	中度	0	0
6	好	1	0	26	中度	1	4
7	好	0	1	27	中度	0	3
8	好	1	3	28	中度	0	9
9	好	1	3	29	中度	1	6
10	好	1	7	30	中度	0	4
11	好	0	1	31	中度	0	3
12	好	0	2	32	受损	1	8
13	轻度	1	5	33	受损	1	2
14	轻度	0	6	34	受损	1	7
15	轻度	1	3	35	受损	0	5
16	轻度	0	1	36	受损	0	4
17	轻度	1	8	37	受损	0	4
18	轻度	1	2	38	受损	1	8
19	轻度	0	5	39	受损	0	8
20	轻度	1	5	40	受损	0	9

a 0,低;1,高。

具有式 7.5 形式的主效应模型是

$$\text{Logit}[P(Y \leqslant j \mid \mathbf{x})] = \alpha_j + \beta_1 x_1 + \beta_2 x_2\text{。}$$

表 7.6 给出了拟合结果。结果变量包括 $J = 4$ 个类别,所以模型具有三个截距项 $\{\alpha_j\}$ 。通常,除了需要计算结果变量的概率外,这些截距没有特别的意义。参数估计给出了所估计的 Logit,进而可以求得关于 $P(Y \leqslant j)$ 、$P(Y > j)$ 或 $P(Y = j)$ 的估计。我们通过生命事件指数取均值 $x_1 = 4.275$ 以及社会经济地位较低($x_2 = 0$)的对象来加以展示。由于 $\hat{\alpha}_1 = -0.282$,结果为"好"的估计概率等于

$$\hat{P}(Y = 1) = \hat{P}(Y \leqslant 1) = \frac{\exp[-0.282 - 0.319(4.275)]}{1 + \exp[-0.282 - 0.319(4.275)]} = 0.16\text{。}$$

图 7.6 区分了社会经济地位的两种情况,将 $\hat{P}(Y > 2)$ 作为生命事件指数的函数进行了绘图。

表 7.6 对表 7.5 数据拟合累积 Logit 模型的输出结果

Score Test for the Proportional odds Assumption

Chi-Square	DF	Pr > ChiSq
2.325 5	4	0.676 1

Parameter	Estimate	Std Error	Like. Ratio 95% Conf Limits		Chi-Square	Pr > Chi Sq
Intercept1	−0.281 9	0.642 3	−1.561 5	0.983 9	0.19	0.660 7
Intercept2	1.212 8	0.660 7	−0.050 7	2.565 6	3.37	0.066 4
Intercept3	2.209 4	0.721 0	0.859 0	3.712 3	9.39	0.002 2
life	−0.318 9	0.121 0	−0.571 8	−0.092 0	6.95	0.008 4
ses	1.111 2	0.610 9	−0.064 1	2.347 1	3.31	0.068 9

译者注——life:生命事件指数;ses:社会经济地位。

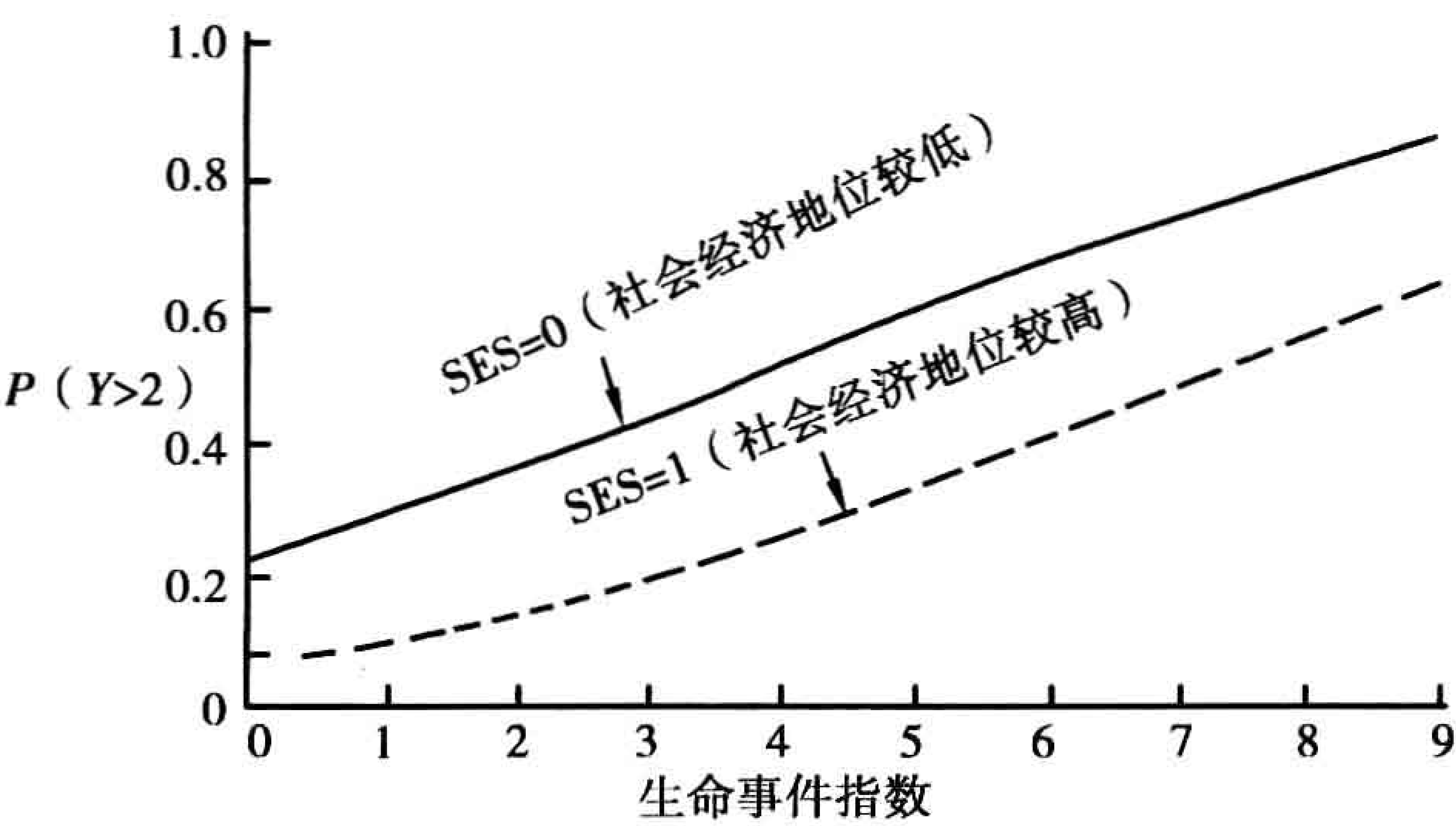


图 7.6 关于表 7.5 中 $P(Y>2)$ 的估计结果

预测变量效应估计的结果 $\hat{\beta}_1=-0.319$ 和 $\hat{\beta}_2=1.111$ 表明,从结果变量的刻度为"好"的一端所开始的累积概率,随着生命事件指数的上升而下降,并且该概率在较高的社会经济地位下也较高。给定生命事件指数的取值,具有较高社会经济地位的研究对象所估计的精神损伤程度低于某一给定水平的发生比是处于较低社会经济地位的估计发生比的 $e^{1.111}=3.0$ 倍。

对效应的描述可以通过比较累积概率而不使用发生比之比来完成,这样更易于理解。对于连续变量,可以通过比较在它们的四分位点处的概率来描述其效应;对于分类变量,我们通过比较它们的不同类别所对应的概率来描述其效应。为了控制某个连续变量,我们将它的值设定为其均值;而对于分类变量,我们则通过给定某个类别来进行控制,如果存在多个类别的话,可以将其设定为表示它的每个虚拟变量的均值。我们再以 $P(Y=1)$ ——结果为"好"为例来对效应描述加以演示。首先,考虑社会经济地位的效应。当生命事件指数取均值 4.275 时,对于社会经济地位较高的一组(即 $x_2=1$),$\hat{P}(Y=1)=0.37$,相应概率在社会经济地位较低的一组(即 $x_2=0$)为0.16。接下来,我们描述生命事件的效应。生命事件指数的第一个和第三个四分位点分别等于 2.0 和 6.5。当社会经济地位为高时,$\hat{P}(Y=1)$ 在这两个四分位点间从0.55 下降为0.22;在社会经济地位为低时,相应的变动是从 0.28 到 0.09(注意,比较在第一个四分位点处的 0.55和0.28 以及在第三个四分位点处的0.22 和0.09 进一步反映了社会经济地位的效应)。对于研究样本,两个预测变量的效应都很显著。

表 7.6 为 SAS 的输出结果,该表还给出了一个关于比例发生比特性的计分检验。它

检验对于每个累积 Logit 是否效应都是相同的,其备择假设为效应不相等。该检验将对 x_1 和 x_2 分别只有一个参数的模型与一个更复杂的模型进行比较,后者中对每个变量都分别具有三个参数,即允许对 $\text{Logit}[P(Y \leqslant 1)]$、$\text{Logit}[P(Y \leqslant 2)]$ 和 $\text{Logit}[P(Y \leqslant 3)]$ 存在不同的效应。在这里,计分检验统计量等于 2.33,自由度为 df = 4,因为更复杂的模型比比例发生比模型多四个参数。结果表明,更复杂的模型并没有显著地改善拟合结果($P = 0.68$)。

7.2.5 更复杂的模型

更复杂的累积 Logit 模型可以通过普通的 logistic 回归来表述。不过,它需要一组截距参数而不是一个。例如,在前面的例子中,允许交互效应的模型的最大似然拟合为

$$\text{Logit}[\hat{P}(Y \leqslant j \mid \mathbf{x})] = \hat{\alpha}_j - 0.420x_1 + 0.371x_2 + 0.181x_1x_2,$$

其中 x_1x_2 的系数具有 SE = 0.238。对于社会经济地位较低一组,所估计的生命事件对累积 Logit 的效应为 -0.420;而对社会经济地位较高一组,该效应则等于 $(-0.420 + 0.181) = -0.239$。生命事件的影响似乎对社会经济地位较低的人更严重一些,但是效应之间的差别并不显著。

这节中的模型应用了比例发生比假定,即对不同的累积 Logit 的效应都是相同的。这样做的一个优点是关于效应的描述和解释比较简单,每个预测变量只需要一个参数。该模型也可以扩展到包括不同效应的情况,将式 7.5 中的 $\boldsymbol{\beta}$ 替代为 $\boldsymbol{\beta}_j$。这意味着,不同 Logits 的曲线之间不再平行。这时,不同累积概率的曲线在某个 $\mathbf{x}$ 值出现交叉。这样的模型违背了累积概率之间应有的排序。

即使上述模型在所观察到的 $\mathbf{x}$ 的取值范围内拟合得更好,考虑到简约性,我们仍然可能选择较简单的模型。其中一种情况是,当效应 $\{\hat{\boldsymbol{\beta}}_j\}$ 在不同 *l*ogit 之间并不存在实质性的差异时,这时比例发生比检验的显著性可能主要反映了很大样本规模 n 的影响。即便当 n 较小时,尽管较简单模型的效应估计是有偏的,它的估计值仍然可能比包括很多参数的更复杂模型对应的估计值具有较小的均方差(MSE)。因此,即使关于比例发生比检验的 P 值很小,仍然不要简单地放弃该模型。

如果从现实意义和统计显著性的角度,比例发生比模型都拟合得不好,我们可以考虑其他的一些策略。这包括:(1)尝试使用一个结果变量曲线不对称的连结函数(如补余双对数连结);(2)在线性预测项中加入更多的项,比如交互项;(3)加入离散参数;(4)允许某些预测变量对每个 Logit 具有不同的效应,但不是所有变量都如此(例如,*部分比例发生比*(*partial proportional odds*));(5)拟合基线类别 Logit 模型,并在解释结果时非正式地利用排序的信息。关于第(4)种方法,参见:Peterson and Harrell(1990)、Stokes et al.(2000, Sec. 15.13),另见 Cox(Cox, 1995)对此的批评。在下一节中,我们将累积 Logit 模型扩展到(1)和(3)的情况。

7.3 定序结果变量:累积连结模型

累积 Logit 模型使用的是 Logit 连结。与单变量的广义线性模型一样,其他的连结函数同样适用。令 G^{-1} 表示一个连结函数,它是连续的累积分布函数 G 的反函数(回顾第 4.2.5节)。*累积连结*(*cumulative link*)模型

$$G^{-1}[P(Y \leqslant j \mid \mathbf{x})] = \alpha_j + \boldsymbol{\beta}'\mathbf{x} \qquad (7.7)$$

将累积概率和线性预测项联系起来。Logit 连结函数 $G^{-1}(u) = \log[u/(1-u)]$ 是标准 logistic 累积分布函数的反函数。

与式 7.5 比例发生比模型一样,对于每个切点,$j=1,\cdots,J-1$,假定式 7.7 中 $\mathbf{x}$ 的效应都相等。在第 7.2.3 节我们显示,当关于潜变量 Y^* 的线性回归具有标准的累积分布函数 G 时,这个假设成立。式 7.7 模型来自于从一组累积分布函数为 $G(y^*-\boldsymbol{\beta}'\mathbf{x})$ 的位置参数中对 Y^* 的离散测量。参数 $\{\alpha_j\}$ 是将潜变量进行标准化处理后的类别切点。在这个意义上,累积连结模型是通过线性预测项 $\boldsymbol{\beta}'\mathbf{x}$ 来描述解释变量对 Y^* 的某种不精确定序测量的效应的回归模型。在线性预测项中,使用 $-\boldsymbol{\beta}$ 而不是 $+\boldsymbol{\beta}$ 只会影响 $\hat{\boldsymbol{\beta}}$ 的符号的变动。大多数软件(如 SAS 和 GENMOD 的 LOGISTIC 模块)在拟合时使用的是 $+\boldsymbol{\beta}$。

7.3.1 累积连结的种类

当 G 是标准正态累积分布函数 Φ 时,对应的是*累积 probit 模型*(*cumulative probit model*)。它将二分 probit 模型(第 6.6 节)扩展到定序结果变量的情况。当 Y^* 服从正态分布时,该模型是理想的选择。在 probit 模型中,可以按照潜变量 Y^* 来对参数进行解释。例如,以模型 $\Phi^{-1}[P(Y\leqslant j)] = \alpha_j - \beta x$ 为例。从第 7.2.3 节可知,由于 $Y^* = \beta x + \varepsilon$,其中 $\varepsilon \sim N(0,1)$ 并具有累积分布函数 Φ,β 可以解释为 x 每增加 1 单位对应着 $E(Y^*)$ 增加 β 个单位。当 ε 不服从 $\sigma=1$ 的标准正态分布时,x 每增加 1 单位对应的是 $E(Y^*)$ 增加 β 个标准差。累积 Logit 模型的拟合结果与累积 probit 模型相似,而且前者的参数解释起来更为简单。

如果 Y^* 服从一个隐含的极值分布,则可以推出以下模型:

$$\log\{-\log[1-P(Y\leqslant j\mid \mathbf{x})]\} = \alpha_j + \boldsymbol{\beta}'\mathbf{x}。$$

在第 6.6 节中,我们介绍了关于二分数据的这一*补余双对数连结*(*complementary log-log link*)。使用该连结的定序模型有时被称为*比例风险*(*proportional hazards*)模型,因为它源自对处理具有离散时间的生存数据的比例风险模型的一般化(Prentice and Gloeckler, 1978)。该模型具有以下特性

$$P(Y>j\mid \mathbf{x}_1) = [P(Y>j\mid \mathbf{x}_2)]^{\exp[\boldsymbol{\beta}'(\mathbf{x}_1-\mathbf{x}_2)]}。$$

在这个连结下,$P(Y\leqslant j)$ 趋近于 1.0 的速度要快于它趋近于 0.0 的速度。在补余双对数连结成立的情况下,如果将结果变量的类别反过来排序,可以考虑使用相应的*双对数连结*(*log-log link*) $\log\{-\log[P(Y\leqslant j)]\}$。

7.3.2 累积连结模型的估计

McCullagh(1980),Thompson 和 Baker(1981)将累积连结模型视为多元广义线性模型。通过累积概率将似然函数表述为式 7.6 的形式,McCullagh 提出了一种进行最大似然估计的 Fisher 计分算法。McCullagh 证明,如果 n 足够大,可以保证似然函数存在一个唯一的最大值。Burridge(1981)和 Pratt(1981)表明,对于许多累积连结模型,包括 Logit, probit 和补余双对数连结,对数似然函数都是凹函数。迭代法通常会迅速收敛于最大似然估计值。

7.3.3 例子:生命表数据

表 7.7 所示为 1981 年美国居民按种族和性别划分的寿命分布。这里,我们用 5 个排序的类别来表示寿命的长度。它背后的累积分布函数在年轻时增长缓慢,但在年老时迅

速上升。这表明,应当使用补余双对数连结。该连结源于对风险率随年龄呈指数上升的假设,也就是极值分布(也即 Gompertz 分布)的情况。

表7.7 1981 年美国居民寿命的分布[a]/%

寿命水平	男性		女性	
	白人	黑人	白人	黑人
0~20	2.4 (2.4)	3.6 (4.4)	1.6 (1.2)	2.7 (2.3)
20~40	3.4 (3.5)	7.5 (6.4)	1.4 (1.9)	2.9 (3.4)
40~50	3.8 (4.4)	8.3 (7.7)	2.2 (2.4)	4.4 (4.3)
50~60	17.5 (16.7)	25.0 (26.1)	9.9 (9.6)	16.3 (16.3)
65 岁以上	72.9 (73.0)	55.6 (55.4)	84.9 (84.9)	73.7 (73.7)

a 括号中的数值为比例风险模型(即补余双对数连结)的拟合值。

来源:Data from *Statistical Abstract of the Unite States*(Washingtong,DC:U.S. Bureau of the Census,1984),p.69。

对于性别 G (1 = 女性;0 = 男性)、种族 R (1 = 黑人;0 = 白人),以及寿命 Y ,表7.7给出了以下模型所拟合的分布:

$$\log\{-\log[1-P(Y\leqslant j\mid G=g,R=r)]\} = \alpha_j+\beta_1 g+\beta_2 r。$$

在这里,拟合优度检验是没有意义的,因为该表所示的是总体而不是样本的分布。模型对数据拟合得很好。它估计的参数值分别为$\beta_1=-0.658$ 和$\beta_2=0.626$ 。拟合的累积分布函数满足

$$P(Y>j\mid G=0,R=r)=[P(Y>j\mid G=1,R=r)]^{\exp(0.658)}。$$

给定种族,男性寿命比某一寿命水平更长的比例等于女性的相应比例的 $\exp(0.658)=1.93$ 次乘方。给定性别,黑人寿命比某一寿命水平更长的比例等于白人的相应比例的 $\exp(0.626)=1.87$ 次乘方。β_1 和β_2 的值表明,白人男性和黑人女性具有相似的分布,而白人女性通常拥有最长的寿命,黑人男性则寿命最短。如果白人女性的寿命比某一给定水平更长的概率等于 π ,白人男性和黑人女性的相应概率大约等于 π^2 ,而黑人男性则约为 π^4。

7.3.4 加入离散效应*

在累积连结模型中,解释变量的不同取值组合之间存在着关于结果变量的*随机排序*(*stochastically ordered*):在任意两个 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 之间,对于所有的 j ,或者 $P(Y\leqslant j\mid \mathbf{x}_1)\leqslant P(Y\leqslant j\mid \mathbf{x}_2)$,或者 $P(Y\leqslant j\mid \mathbf{x}_1)\geqslant P(Y\leqslant j\mid \mathbf{x}_2)$ 。图7.7a 展示了在 $\mathbf{x}$ 的两个不同取值处,结果变量对应的潜变量的连续密度函数和累积分布函数。当数据违背了这种随机排序并且模型的拟合结果不好时,通常是因为结果变量的离散程度随着 $\mathbf{x}$ 的变动而发生变动。例如,结果变量可能倾向于集中在同一位置,但是它在 $\mathbf{x}_1$ 处的离散程度比在 $\mathbf{x}_2$ 处更高。这时,可能会出现对于较小的j,存在 $P(Y\leqslant j\mid \mathbf{x}_1)>P(Y\leqslant j\mid \mathbf{x}_2)$,而对于较大的j,存在 $P(Y\leqslant j\mid \mathbf{x}_1)<P(Y\leqslant j\mid \mathbf{x}_2)$ 。换句话说,结果变量在 $\mathbf{x}_1$ 处的分布比在 $\mathbf{x}_2$ 处更集中于一些极端的类别。图7.7b 描述了这种情况下的隐含连续分布。

在累积连结模型中加入离散效应,我们得到模型

$$G^{-1}[P(Y\leqslant j\mid \mathbf{x})]=\frac{\alpha_j+\boldsymbol{\beta}'\mathbf{x}}{\exp(\boldsymbol{\gamma}'\mathbf{x})}。\tag{7.8}$$

(同样,大家可以将 + 替代为 - ,以使它能更好地切合所隐含的连续变量的位置-规模族(location-scale family)分布)。分母包括了规模参数(scale parameters)$\boldsymbol{\gamma}$,用来描述离散度随着 $\mathbf{x}$ 的变动。普通的模型式7.7 是上述模型中当 $\boldsymbol{\gamma}=0$ 时的一个特例,否则,当$\boldsymbol{\gamma}'\mathbf{x}>0$

时，累积概率倾向于相互收缩在一起。这导致位于边界的类别对应的概率变高，并且总的离散度变大。当 $\boldsymbol{\gamma}'\mathbf{x}<0$ 时，累积概率倾向于相互远离（导致较小的离散度）。

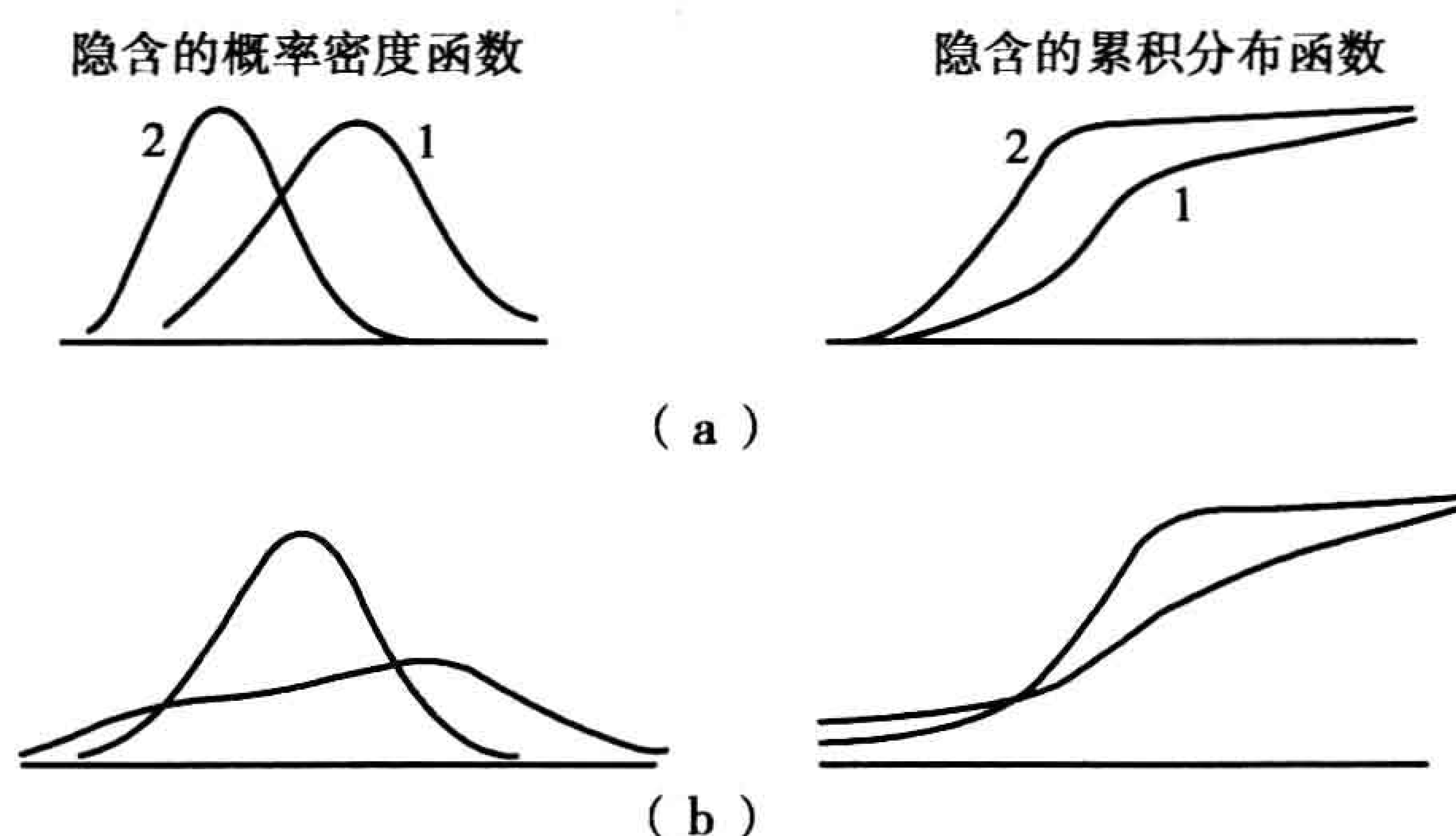


图 7.7 (a)分布 1 随机高于分布 2；(b)分布之间不存在随机排序

具体来说，我们利用这个模型比较在一个定序尺度上的两个组别。假定 x 是一个虚拟变量，其中 $x=1$ 表示第 1 组。在累积 Logit 的形式下，式 7.8 模型为

$$\text{Logit}[P(Y\leqslant j)]=\alpha_j,\quad x=0,$$

$$\text{Logit}[P(Y\leqslant j)]=(\alpha_j+\beta)/\exp(\gamma),\quad x=1。$$

在 $\gamma=0$ 的情况下，我们得到通常的模型，其中 β 表示将一个 $2\times J$ 表格进行合并得到的所有可能的 2×2 表格中共同累积对数发生比之比的位置移动。当 $\gamma\neq0$ 时，两个组别间的 Logit 以及相应的累积发生比之比，随着 j 的变化而变化。当 $\gamma>0$ 时，结果变量在 $x=1$ 时比在 $x=0$ 时更分散。关于该模型的拟合及示例，参见：*Cox*（1995）、*McCullagh*（1980）。

7.4 关于定序结果变量的其他模型*

定序结果变量模型并不一定通过累积概率来表示。在本节，我们讨论其他形式的 Logit 模型以及类似于普通回归的较简单模型。

7.4.1 相邻类别 Logit

相邻类别 Logit（adjacent-categories Logits）为

$$\text{Logit}[P(Y=j\mid Y=j\text{ 或 }j+1)]=\log\frac{\pi_j}{\pi_{j+1}},\quad j=1,\cdots,J-1。\tag{7.9}$$

这些 Logit 是与基线类别 Logit 等价的一组基本集。它们之间的关系是

$$\log\frac{\pi_j}{\pi_J}=\log\frac{\pi_j}{\pi_{j+1}}+\log\frac{\pi_{j+1}}{\pi_{j+2}}+\cdots+\log\frac{\pi_{J-1}}{\pi_J},\tag{7.10}$$

以及

$$\log\frac{\pi_j}{\pi_{j+1}}=\log\frac{\pi_j}{\pi_J}-\log\frac{\pi_{j+1}}{\pi_J},\quad j=1,\cdots,J-1。$$

以上任一组 Logit 都可以决定结果变量类别两两配对之间的所有 $\binom{J}{2}$ 个 Logit。

利用相邻类别 Logit 的模型也可以表述为相应的基线类别 Logit 模型。例如，考虑以下相邻类别 Logit 模型：

$$\log\frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})}=\alpha_j+\boldsymbol{\beta}'\mathbf{x},\quad j=1,\cdots,J-1,\tag{7.11}$$

它具有共同的效应 $\boldsymbol{\beta}$。在其中加入类似于式 7.10 的 $(J-j)$ 项，与之等价的基线类别 Logit 模型是

$$\log\frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} = \sum_{k=j}^{J-1}\alpha_k + \boldsymbol{\beta}'(J-j)\mathbf{x}$$

$$= \alpha_j^* + \boldsymbol{\beta}'\mathbf{u}_j,\quad j = 1,\cdots,J-1,$$

其中 $\mathbf{u}_j = (J-j)\mathbf{x}$。相邻类别 Logit 模型对应着一个将模型矩阵进行调整后的基线类别 Logit 模型，但是每个预测变量仍只对应一个参数。利用有关统计软件，大家可以通过拟合等价的基线类别 Logit 模型来拟合式 7.11 模型。

相邻类别 Logit 的构建反映了 Y 的类别之间的排序特征。为了在模型简约性方面对此加以利用，需要对线性预测项做适当的设定。例如，如果一个解释变量对每个 Logit 的效应都相似，那么使用一个而不是 $(J-1)$ 个参数来描述该效应更具优势。在比例发生比模型中，相邻类别 Logit 模型（式 7.11）与使用累积 Logit 的模型（式 7.5）拟合情况相似。它们都隐含着在预测变量的不同取值组合上，Y 的分布是随机排序的。

关于模型的选择不应当过于依赖拟合优度，而是要考虑对不同效应形式的偏好，是偏好相邻类别 Logit 中针对结果变量的每个类别的效应，还是累积 Logit 中强调对所有类别的整个刻度或一个隐含的潜变量的效应。由于在累积 Logit 模型中，效应是针对整个刻度而言的，它的值通常会较大。不过，在两种模型中，估计值和标准误的比通常都很相似。累积 Logit 模型的一个优点在于，它对效应的估计基本不会随着结果变量的类别数量和切点位置的变化而变化。这一点对于相邻类别 Logit 并不成立。

7.4.2　例子：工作满意度

表 7.8 所示为按照性别进行分层后，美国黑人工作满意度（Y）与收入的关系。简便起见，我们使用赋值（1,2,3,4）来表示收入。对于收入 x 和性别 g（1 = 女性，0 = 男性），考虑以下模型：

$$\text{Logit}(\pi_j/\pi_{j+1}) = \alpha_j + \beta_1 x + \beta_2 g,\quad j = 1,2,3。$$

它描述的是对工作非常不满意而不是有点儿满意的发生比、有点儿满意而不是基本满意的发生比，以及基本满意而不是非常满意的发生比。

表 7.8　控制性别后的工作满意度与收入

性　别	收　入	工作满意度			
		非常不满意	有点儿满意	比较满意	非常满意
女　性	<5 000	1	3	11	2
	5 000 ~ 15 000	2	3	17	3
	15 000 ~ 25 000	0	1	8	5
	>25 000	0	2	4	2
男　性	<5 000	1	1	2	1
	5 000 ~ 15 000	0	3	5	1
	15 000 ~ 25 000	0	0	7	3
	>25 000	0	1	9	6

来源：1991 年美国综合社会调查，美国民意研究中心（National Opinion Reserch Center）。

这个模型等价于以下基线类别 Logit 模型：

$$\log(\pi_j/\pi_4) = \alpha_j^* + \beta_1(4-j)x + \beta_2(4-j)g,\quad j = 1,2,3。$$

在关于 $\log(\pi_1/\pi_4)$ 的方程中，这个模型中第一个预测项的值被设定为等于 $3x$，在关于

$\log(\pi_2/\pi_4)$ 的方程中它等于 $2x$，以及在 $\log(\pi_3/\pi_4)$ 的方程中它等于 x。有些软件（例如，SAS 的 PROC CATMOD，参见附表 A.12）允许在预测项的给定取值处输入一行关于每个基线类别 Logit 的模型矩阵。这时，在拟合了限定对每个 Logit 的效应相同的基线类别 Logit 模型后，估计的回归参数就是对相邻类别 Logit 模型参数的最大似然估计。最大似然拟合的结果为 $\hat{\beta}_1 = -0.389$（SE = 0.155）和 $\hat{\beta}_2 = 0.045$（SE = 0.314）。按照这种参数化方式，$\hat{\beta}_1 < 0$ 表示，随着收入的上升，较低工作满意度的发生比会下降。在给定性别的情况下，收入每上升一个类别，所估计的结果变量处在两个相邻的类别中较低一个的发生比需要乘以 $\exp(-0.389) = 0.68$。这个模型通过 5 个参数描述了 24 个 Logit（收入和性别的每个取值组合各对应三个）。模型的偏离度 $G^2 = 12.6$，df = 19。因而，这个将收入效应表示为线性趋势，而且不包括收入与性别的交互项的模型看起来是充分的。

通过拟合累积 Logit 模型，也可以得出相似的结论。它具有偏离度 $G^2 = 13.3$，df = 19。该模型中收入的效应更大一些（$\hat{\beta}_1 = -0.51$，SE = 0.20），因为它所对应的是结果变量的整个刻度而不仅仅是相邻的类别。不过，两种模型下的显著性水平很相似，都有 $\hat{\beta}_1/\text{SE} \approx -2.5$。

7.4.3 连续比 Logit

连续比（*continuation-ratio*）Logit 的定义为

$$\log \frac{\pi_j}{\pi_{j+1} + \cdots + \pi_J}, \quad j = 1, \cdots, J-1 \tag{7.12}$$

或者

$$\log \frac{\pi_{j+1}}{\pi_1 + \cdots + \pi_j}, \quad j = 1, \cdots, J-1。 \tag{7.13}$$

当结果变量发生的过程存在序列特征时，比如各个年龄段的存活情况，连续比 Logit 模型非常有意义（如 Tutz，1991）。令 $\omega_j = P(Y = j \mid Y \geqslant j)$，在包括解释变量的情况下，

$$\omega_j(\mathbf{x}) = \frac{\pi_j(\mathbf{x})}{\pi_j(\mathbf{x}) + \cdots + \pi_J(\mathbf{x})}, \quad j = 1, \cdots, J-1。 \tag{7.14}$$

式 7.12 连续比 Logit 是关于这些条件概率的普通 Logit，即 $\log[\omega_j(\mathbf{x})/(1-\omega_j(\mathbf{x}))]$。

在 $\mathbf{x}$ 的第 i 个取值 $\mathbf{x}_i$ 处，令 $\{y_{ij}, j = 1, \cdots, J\}$ 表示结果变量的计数，并有 $n_i = \sum_j y_{ij}$。当 $n_i = 1$ 时，y_{ij} 表示结果变量的取值是否落在第 j 个类别，这与第 7.1.4 节相同。令 $b(n, y; \omega)$ 表示在 n 次试验中成功次数为 y 的二项分布概率，每次试验的参数为 ω。通过将多项分布概率 $(y_{i1}, \cdots, y_{iJ})$ 表示为 $p(y_{i1})p(y_{i2} \mid y_{i1}) \cdots p(y_{iJ} \mid y_{i1}, \cdots, y_{i,J-1})$ 的形式，可以对多项分布密度函数进行以下因式分解：

$$\begin{aligned} & b[n_i, y_{i1}; \omega_1(\mathbf{x}_i)]\, b[n_i - y_{i1}, y_{i2}; \omega_2(\mathbf{x}_i)] \cdots \\ & b[n_i - y_{i1} - \cdots - y_{i,J-2}, y_{i,J-1}; \omega_{J-1}(\mathbf{x}_i)]。 \end{aligned} \tag{7.15}$$

整个似然函数是对在不同的 $\mathbf{x}_i$ 值处的多项分布密度函数的乘积。因此，对数似然函数等于对包括不同 ω_j 的各项的求和。对于任意的 $j \neq k$，当在模型设定中关于 Logit（ω_j）的参数与 Logit（ω_k）的参数不同时，将每个单独的项分别进行最大化同时也使得整个对数似然函数取最大值。这样，对不同的连续比 Logit 分别拟合模型给出的结果与同时拟合整个模型的结果是一致的。相应地，将 $J-1$ 个模型的 G^2 统计量进行求和给出了一个与同时拟合所有 $J-1$ 个模型相同的总体拟合优度统计量。

因为这些 Logit 针对的是一个二分结果变量，其中一类是对原有类别的合并，可以应用

有关拟合二分 Logit 模型的方法来分别拟合这些模型。类似的结论也适用于式 7.13 连续比 Logit,尽管这些 Logit 以及后续分析的结果可能与式 7.12 不同。有时,对每个 Logit 都具有相同效应的较简单模型可能更加合理(McCullagh and Nelder,1989,p. 164;Tutz 1991)。

7.4.4 关于受孕老鼠的发育毒性研究

我们通过表 7.9 中一项有关发育毒性(developmental toxicity)的研究来展示连续比 Logit 模型。这种研究利用老鼠进行实验,来检验某些危害性物质对胎儿发育的可能影响。该研究所考察的物质为二乙二醇二甲醚(Diethylene glycol dimethyl ether, diEGdiME),它是一种用于保护性涂层如喷漆和金属涂层等的工业溶剂。

表 7.9 一项利用受孕老鼠进行发育毒性研究的结果[a]

浓度	结果		
(毫克/千克每天)	死亡	畸形	正常
0(控制组)	15	1	281
62.5	17	0	225
125	22	7	283
250	38	59	202
500	144	132	9

a 基于 C. J. Price et al. (C. J. Price et al. *Fund. Appl. Toxicol.* 8:115-116(1987))的研究结果。感谢 Louise Ryan 给我提供了该数据。

这项研究给受孕老鼠饮用的蒸馏水中加入了 diEGdiME。这些老鼠在受孕初期分别服用了 5 种不同浓度水平的 diEGdiME,周期为 10 天。两天后,对老鼠的子宫内含物进行了缺陷检查。每个胚胎存在三种可能的结果(死胎,畸形,正常)。这些结果具有排序性,其中死胎是最差的结果。我们通过连续比 Logit 对出现死胎的概率 π_1 以及给定胚胎还活着的情况下发生畸形的条件概率 $\pi_2/(\pi_2+\pi_3)$ 来进行模型分析。

我们拟合了连续比 Logit 模型

$$\log\frac{\pi_1(x_i)}{\pi_2(x_i)+\pi_3(x_i)}=\alpha_1+\beta_1 x_i,\quad \log\frac{\pi_2(x_i)}{\pi_3(x_i)}=\alpha_2+\beta_2 x_i,$$

这里,利用赋值 $\{0,62.5,125,250,500\}$ 来表示浓度水平 x_i。最大似然估计的结果为 $\hat{\beta}_1=0.0064(SE=0.0004)$,$\hat{\beta}_2=0.0174(SE=0.0012)$。在每种情况下,随着浓度的增加,较差的结果发生的可能性也增大。例如,在给定某个胚胎仍活着的情况下,随着 diEGdiME浓度每增加 100 个单位,所估计的出现畸形的发生比要乘以 $\exp(1.74)=5.7$。模型的似然比拟合统计量为:对于 $j=1, G^2=5.78$;对于 $j=2, G^2=6.06$,每个模型都是基于自由度 df=3。将二者相加,$G^2=11.84$(或相似地,$X^2=9.76$),自由度 df=6,这些结果给出了模型的总体拟合情况。

这个分析将不同胚胎的受孕结果视为相互独立的、同等的观测值。事实上,每个受孕老鼠都怀有一窝胚胎,并且同一窝的不同胚胎之间可能存在着统计上的相依性(statistical dependence)。在某个给定的浓度水平上,不同窝的胚胎所具有的结果变量的概率可能也不同。在各窝之间的种种异质性(例如,不同受孕老鼠之间的体征差异)会导致这些概率之间存在某种变动。统计相依性或者异质性概率都会违背二项分布的假定,并导致过度离散的问题。在某个给定的浓度水平上,一窝胚胎中死亡的数目在受孕老鼠之间的变动可能会比如果这些计数是独立的、服从同一二项分布的情况大。利用该数据拟合的总体 G^2 显示出了某种程度的拟合不足($P=0.07$),这很可能反映了由上述因素

导致的过度离散的影响,而不是由于所选用的结果变量曲线不当所致。

为了处理过度离散的问题,我们可以通过类似然法对标准误进行调整(第 4.7 节)。这种方法将标准误乘以 $\sqrt{X^2/\mathrm{df}} = \sqrt{9.76/6} = 1.28$ 。对于每个 Logit,调整后仍然存在强有力的证据表明 $\beta_j > 0$ 。在第 12 章和第 13 章中,我们将介绍处理各窝中胚胎的群组问题(clustering)的其他方法。

7.4.5 关于定序结果变量的平均结果变量模型

现在我们介绍一个与连续结果变量的普通回归相类似的模型。对于赋值 $v_1 \leq v_2 \leq \cdots \leq v_J$,令

$$M(\mathbf{x}) = \sum_j v_j \pi_j(\mathbf{x})$$

表示结果变量的平均值。模型

$$M(\mathbf{x}) = \alpha + \boldsymbol{\beta}'\mathbf{x} \tag{7.16}$$

假定在该平均值与解释变量之间存在线性关系。当 $J = 2$ 时,它相当于线性概率模型(第 4.2.1 节)。当 $J > 2$ 时,该模型并不从结构上设定结果变量的发生概率,而是仅仅描述它的平均值取决于 $\mathbf{x}$。

假定在不同的 $\mathbf{x}_i$ 处,观测值是独立的多项分布样本,Bhapkar(1968)、Grizzle 等(1969),以及 Williams 和 Grizzle(1972)介绍了有关平均结果变量模型(mean response models)的加权最小二乘法(weighted least squares,WLS)拟合。将在第 15.1 节介绍的加权最小二乘法适用于所有解释变量都是分类变量的情况。无论解释变量是分类还是连续变量,都可以应用最大似然法,最大化多项分布似然函数之间的乘积,Haber(1985)和 Lipsitz(1992)提出了对包括平均结果变量模型在内的一组模型进行最大似然拟合的算法。这种算法比较复杂,因为多项分布似然函数中的概率并不直接就是式 7.16 模型中的参数的函数,但是,可以通过专门的软件来处理这一问题(参见附录 A)。

7.4.6 例子:再论工作满意度

我们通过表 7.8 来展示平均结果变量模型,利用收入 x 和性别 g (1 = 女性, 0 = 男性),对 Y = 工作满意度的平均值进行模型分析。为简化起见,我们使用 (1,2,3,4) 作为对工作满意度和收入的赋值。模型的最大似然拟合结果为

$$\hat{M} = 2.59 + 0.181x - 0.030g,$$

其中,关于收入的标准误为 SE = 0.069,关于性别的标准误为 SE = 0.145。在给定性别的情况下,收入每上升一个类别,平均工作满意度估计上升大约 0.2 个类别。尽管模型结果显示收入存在显著的正效应(例如,沃尔德统计量 $(0.181/0.069)^2 = 6.8$,df = 1, $P = 0.009$),但是效应本身并不大。在最高收入水平上所估计的工作满意度只比在最低收入水平上平均高大约半个类别,因为 $3(0.181) = 0.54$ 。运用加权最小二乘法得出的结果与之相似,它所估计的收入效应为 0.182,SE = 0.068(附表 A.12 显示了 SAS 中 CATMOD 给出的结果)。

检验模型拟合优度的偏离度等于 5.1。由于平均值发生在 8 个收入与性别的取值结合点,而模型本身具有 3 个参数,所以残差自由度 df = 5。拟合结果看上去是充分的。

7.4.7 平均结果变量模型的优缺点

如果定序变量的类别在本质上反映了对一个内在的连续变量的粗略测量,那么将其

作为一种连续变量来处理是有道理的。平均结果变量模型具有与普通回归非常相似的优点。

当 $J = 2$ 时,我们在第 4.2.1 节指出,由于概率本身的限定为 (0,1) ,线性概率模型具有一种结构性缺陷。这里也存在相似的问题,因为线性模型的预测值可能会超出对结果变量平均值的赋值范围。当 J 较大并且在所关注的解释变量的取值范围内结果变量存在较为合理的离散度时,这种问题出现的可能性会小一些。对于定序变量来说,其背后存在一个隐含的潜变量的思想比严格的二分变量更合理一些,因而,在这里这个问题没那么严重。

与 Logit 模型不同,平均结果变量模型并不唯一决定单元格概率,因此,平均结果变量模型不需要设定比如随机排序等结构性特性。它不像针对概率的模型那样完全体现结果变量的分类结构,并且独立性等条件也不等于它的某个特例,但是,它对数据的描述比累积连结模型的发生比之比或总体指标更为简单。随着 J 的增大,平均结果变量模型直接对应于普通回归模型。在 J 较大的情况下,它是对如果我们能够对 Y 进行连续尺度的测量并对其进行普通回归分析的一种简单近似。

7.5 $I \times J \times K$ 表格中的条件独立性检验*

在第 6.3.2 节中,我们介绍了关于 $2 \times 2 \times K$ 表格的条件独立性的 Cochran-Mantel-Haenszel(CMH)检验。本节介绍由多类别结果变量形成的 $I \times J \times K$ 表格的相应检验。通过比较设定 XY 条件独立性的模型与 XY 具有相依性的模型的拟合结果,就得出相应的似然比检验。还有一种方法是对 CMH 统计量的扩展,它也是某些模型的计分统计量。

7.5.1 利用多项分布模型检验条件独立性

将 Z 视为一个定类的控制变量,我们讨论 (Y,X) 分别为(定序,定序)、(定序,定类)、(定类、定序),以及(定类,定类)的四种情况。对于 Y 为定序变量的情况,我们使用累积 Logit 模型,但是使用其他的定序连结也会得到相似的结果。正如我们在第 6.3.2 节所指出的,当分表之间的 XY 关联相似时,基于同质性关联模型的检验统计量可以提高统计效能。

1. *Y 定序,X 定序*。令 $\{x_i\}$ 表示有序的赋值。模型

$$\text{Logit}[P(Y \leqslant j \mid X = i, Z = k)] = \alpha_j + \beta x_i + \beta_k^Z \qquad (7.17)$$

在每个分表中,X 的效应具有相同的线性趋势。对此,XY 条件独立性对应于 $H_0: \beta = 0$ 。关于 H_0 的似然比、计分或沃尔德统计量提供了以线性趋势作为备择假设的 df = 1 的大样本卡方检验。

2. *Y 定序,X 定类*。一种关于条件独立性的表述是将 X 视为一个定类的因子

$$\text{Logit}[P(Y \leqslant j) \mid X = i, Z = k] = \alpha_j + \beta_i + \beta_k^Z ,$$

其中限定条件为 $\beta_I = 0$ 。对此模型,XY 条件独立性由 $H_0: \beta_1 = \cdots = \beta_I$ 来表示。大样本卡方检验的自由度为 df = $I - 1$。

3. *Y 定类,X 定序*。当 Y 是定类变量时,相应的检验使用基线类别 Logit 模型。反映 XY 条件独立性的模型为

$$\log \frac{P(Y = j \mid X = i, Z = k)}{P(Y = J \mid X = i, Z = k)} = \alpha_{jk}。 \qquad (7.18)$$

对于有序的赋值 $\{x_i\}$,关于在每个分表中存在相同的线性趋势的检验将该模型与

以下模型进行比较：

$$\log \frac{P(Y=j \mid X=i, Z=k)}{P(Y=J \mid X=i, Z=k)} = \alpha_{jk} + \beta_j x_i。$$

条件独立性意味着 $H_0: \beta_1 = \cdots = \beta_{J-1} = 0$。大样本卡方检验的自由度为 $\mathrm{df} = J-1$。

4. Y 定类，X 定类。将 X 视为一个定类的因子，XY 条件独立性可表示为

$$\log \frac{P(Y=j \mid X=i, Z=k)}{P(Y=J \mid X=i, Z=k)} = \alpha_{jk} + \beta_{ij}, \tag{7.19}$$

其中对每个 j 设限定条件如 $\beta_{Ij} = 0$。对于每个 j，X 和 Z 具有 $\alpha_k + \beta_i$ 形式的可加效应。条件独立性等于 $H_0: \beta_{1j} = \cdots = \beta_{Ij}, j = 1, \cdots, J-1$。大样本卡方检验的自由度为 $\mathrm{df} = (I-1)(J-1)$。

表 7.10 对这四种情况进行了总结。当模型至少在某一方面偏离条件独立性时，上述检验都表现很好。这并不意味着为了使用这些检验，大家必须检验模型的拟合（参见第 6.3.2 节末尾的评论）。

有时，K 个分表之间的关联可能存在很大的差异。当 Z 是一个定序变量时，允许对数发生比之比按照 Z 的类别呈线性变动的方法可能会很有用。例如，当 Z = 研究对象的年龄时，历险因素 X（如吸烟频繁度）与结果变量 Y（如心脏病的严重程度）之间的关联可能会随着 Z 的上升而上升。当 Z 是定类变量时，可以通过对比条件独立性模型与允许在 Z 的每个类别上对应不同效应参数的更一般化的模型来进行检验。但是，允许在 Z 的每个取值上存在不同效应会导致检验的自由度上升 K 倍，这会影响检验的效能。

表 7.10　关于检验条件独立性的模型的总结

Y-X	模　型	条件独立性	df
定序-定序	$\mathrm{Logit}[P(Y \leqslant j)] = \alpha_j + \beta x_i + \beta_k^Z$	$\beta = 0$	1
定序-定类	$\mathrm{Logit}[P(Y \leqslant j)] = \alpha_j + \beta_i + \beta_k^Z$	$\beta_1 = \cdots = \beta_I$	$I-1$
定类-定序	$\log\left[\frac{P(Y=j)}{P(Y=J)} = \alpha_{jk} + \beta_j x_i\right]$	$\beta_1 = \cdots = \beta_{J-1} = 0$	$J-1$
定类-定类	$\log\left[\frac{P(Y=j)}{P(Y=J)} = \alpha_{jk} + \beta_{ij}\right]$	所有 $\beta_{ij} = 0$	$(I-1)(J-1)$

7.5.2　例子：再论工作满意度

现在，我们重新来看工作满意度的数据（表 7.8）。表 7.11 总结了几个模型的拟合结果。模型将收入作为定序的预测变量，使用的赋值为 {3,10,20,35}，即以千美元计的每个类别的大致中点。在控制了性别后，每个似然比检验将给定模型与删除了收入效应的模型进行比较。

表 7.11　关于表 7.8 条件独立性的模型似然比检验的总结

满意度	收　入	G^2 拟合	df	检验统计量	df	P 值
定序	定序	13.95	19	5.7	1	0.017
	定类	10.51	17	9.1	3	0.028
	模型未包括	19.62	20	—	—	—
定类	定序	11.74	15	7.6	3	0.054
	定类	7.09	9	12.3	9	0.198
	模型未包括	19.37	18	—	—	—

通过式 7.17 累积 Logit 模型进行条件独立性检验的似然比统计量为 19.62 − 13.95 =

5.7,自由度为 df = 20 - 19 = 1,结果表明存在很显著的收入效应。而将其中一个或两个变量都视为定类变量的模型给出的证据则没有这么强。将检验的备择假设限定为线性趋势,会得出较小的 P 值。但是,正如在第 7.4.2 节和第 7.4.6 节所说的,通过参数估计我们能获得比显著性检验更多的信息。

7.5.3 关于 $I \times J \times K$ 表格的广义 Cochran-Mantel-Haenszel 检验

Birch(1965)、Landis 等(1978),以及 Mantel 和 Byar(1978)对 CMH 统计量进行了扩展(第 6.3.2 节)。这个检验将 X 和 Y 视为对称的,所以相应的三种情况是:将二者都视为定类变量,都视为定序变量,或者一个定序、一个定类变量。以行总计和列总计为条件,每个层具有 $(I-1)(J-1)$ 个非冗余的单元格计数。令

$$\mathbf{n}_k = (n_{11k}, n_{12k}, \cdots, n_{1,J-1,k}, \cdots, n_{I-1,J-1,k})'。$$

在 H_0:条件独立性下,令 $\boldsymbol{\mu}_k = E(\mathbf{n}_k)$,即

$$\boldsymbol{\mu}_k = (n_{1+k}n_{+1k}, n_{1+k}n_{+2k}, \cdots, n_{I-1,+,k}n_{+,J-1,k})'/n_{++k}。$$

令 $\mathbf{V}_k$ 表示 $\mathbf{n}_k$ 的零分布协方差矩阵(null covariance matrix),其中

$$\mathrm{cov}(n_{ijk}, n_{i'j'k}) = \frac{n_{i+k}(\delta_{ii'}n_{++k} - n_{i'+k})n_{+jk}(\delta_{jj'}n_{++k} - n_{+j'k})}{n_{++k}^2(n_{++k} - 1)},$$

这里,当 $a = b$ 时 $\delta_{ab} = 1$,否则 $\delta_{ab} = 0$。

最一般性的统计量将行和列都视为不具有排序特征的定类变量。对 K 个层进行求和,令

$$\mathbf{n} = \sum \mathbf{n}_k, \quad \boldsymbol{\mu} = \sum \boldsymbol{\mu}_k, \quad \mathbf{V} = \sum \mathbf{V}_k。$$

在 X 和 Y 都是定类变量的情况下,广义 CMH 统计量等于

$$\mathrm{CMH} = (\mathbf{n} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{n} - \boldsymbol{\mu})。 \tag{7.20}$$

它的大样本卡方分布具有自由度 df $= (I-1)(J-1)$,其中自由度的大小等于比较式 7.18和式 7.19 Logit 模型的统计量所对应的自由度。两个统计量都对发现在每个层中是否存在相似的条件关联很灵敏。在层数 $K = 1$ 以及观测值个数为 n 的情况下,CMH = $[(n-1)/n]X^2$,其中 X^2 等于式 3.10 皮尔逊统计量。

Mantel(1963)介绍了一个 X 和 Y 都是定序变量情况下的广义检验。利用有序的赋值 $\{u_i\}$ 和 $\{v_j\}$,该检验对于每层具有同方向的相关关系非常灵敏。如果在每个层内,$T_k = \sum_i \sum_j u_i v_j n_{ijk}$ 都超过了它的零分布期望值,那么检验就表明存在正向的趋势。给定每层的边际总计,在条件独立性下,

$$E(T_k) = \left[\sum_i u_i n_{i+k}\right]\left[\sum_j v_j n_{+jk}\right]/n_{++k},$$

$$\mathrm{var}(T_k) = \frac{1}{n_{++k} - 1}\left[\sum_i u_i^2 n_{i+k} - \frac{(\sum_i u_i n_{i+k})^2}{n_{++k}}\right] \times \left[\sum_j v_j^2 n_{+jk} - \frac{(\sum_j v_j n_{+jk})^2}{n_{++k}}\right]。$$

统计量 $[T_k - E(T_k)]/[\mathrm{var}(T_k)]^{1/2}$ 等于第 k 层中 X 和 Y 的相关系数再乘以 $\sqrt{n_{++k} - 1}$。为了对 K 个层的情况进行总结,Mantel(1963)提出了

$$M^2 = \frac{\left\{\sum_k\left[\sum_i \sum_j u_i v_j n_{ijk} - E\left(\sum_i \sum_j u_i v_j n_{ijk}\right)\right]\right\}^2}{\sum_k \mathrm{var}\left(\sum_i \sum_j u_i v_j n_{ijk}\right)}。 \tag{7.21}$$

它近似于χ_1^2的零分布,与在式7.17定序模型中检验$H_0: \beta = 0$相同。当$K = 1$时,它就是式3.15中M^2统计量。

Landis等(1978)提出了一个更具一般性的统计量,式7.20和式7.21都是它的特例。该统计量也可以将X作为定类而将Y作为定序变量,比较I行的平均值与它们的零分布期望值并加以总结,它的自由度为$df = I - 1$(参见注解7.7)。

7.5.4 例子:再论工作满意度

表7.12给出了对表7.8的数据进行广义CMH检验的结果。将变量作为定序变量的统计量对收入所使用的赋值为{3,10,20,35},对工作满意度所使用的赋值为{1,3,4,5}(附表A.12显示了SAS的PROC FREQ所给出的相应结果,但是其中所使用的赋值有所不同)。

表7.12 对工作满意度和收入数据进行广义Cochran-Mantel-Haenszel检验的结果

Summary Statistics for income by satisf

Controlling for gender

Cochran-Mantel-Haenszel Statistics(Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.156 3	0.013 1
2	Row Mean Scores Differ	3	9.034 2	0.028 8
3	General Association	9	10.200 1	0.334 5

*一般性关联(general association)*假设将X和Y作为定类变量并应用了式7.20。它对在Z的每个取值上的任何相似的关联都很灵敏。*行平均计分不等(row mean scores differ)*假设将行作为定类变量而将列作为定序变量。它对在Z每个的取值上I个关于Y的行平均计分的相似变动很敏感。最后,*非零相关(nonzero correlation)*假设将X和Y作为定序变量并应用了式7.21,它对在Z的每个取值上相似的线性趋势很敏感。与表7.11所总结的基于模型的分析结果一致,使用$df = 1$的定序检验所给出的证据更强。

7.5.5 多项Logit模型的计分检验

与第7.5.1节中利用多项Logit模型所进行的检验相比,广义CMH检验似乎是不基于模型的。不过,它们之间存在着非常紧密的联系。对于各种多项Logit模型来说,广义CMH检验就是相应的计分检验。

将X和Y作为定类变量的广义CMH检验(式7.20)是在Logit模型(式7.19)中关于$(I-1)(J-1)$个参数$\{\beta_{ij}\}$等于0的计分检验。将X和Y作为定序变量而应用M^2所进行的广义CMH检验是在式7.17模型中关于$\beta = 0$的计分检验。对于累积Logit模型,模型中$\{x_i\}$的赋值与M^2相一致,并且M^2中$\{v_j\}$的赋值为平均秩赋值(average rank scores)。对于式7.17形式的相邻类别Logit模型,M^2中$\{v_j\}$的赋值是任意的等距赋值。

当各层的样本都很大时,广义CMH检验的结果与比较相应模型的似然比检验结果相似。基于模型的方法的一个好处是它提供了对效应的估计,而广义CMH检验的优点则在于,在K随着n的增加而增加所导致的稀疏渐近特性下,该检验也能保证很好的效果。第6.3.4节的结论在这里依然成立。

7.5.6 条件独立性的精确检验

原则上,可以使用广义CMH统计量进行条件独立性的精确检验,对第6.7.5节中有关

$2 \times 2 \times K$ 表格的情况加以扩展。为了消除冗余参数，在每层中都以行总计和列总计为条件。各层中计数的分布服从多元超几何分布（第 3.5.7 节），由此推导出一个所关注统计量的精确条件分布。检验的 P 值等于在与观察到的层边际相同的表格中检验统计量至少与观察到的情况一样大的概率（参见：Birch，1965；Kim and Agresti，1997；Mehta et al.，1988）。

7.6 离散选择多项 Logit 模型*

关于多项 Logit 模型的一个重要应用就是，分析解释变量对从一组离散的选项中做出选择的效应。这样的例子包括，上班时对交通工具的选择（自驾车，公车，地铁，走路，自行车），住房选择（自有独立住宅，自有公寓，租房子），主要的购物场所（商业区，购物中心，邮购，网上购物），或者产品的品牌等。结果变量为一组离散选项的模型被称为*离散选择模型*（*discrete-choice models*）。

7.6.1 离散选择模型

在许多离散选择的应用中，对于结果变量的不同选项，解释变量可以取不同的值。作为选择交通方式的预测变量，到达目的地所需要的花费和时间在不同的交通工具下是不一样的。作为选取产品品牌的预测变量，价格会随着选择结果的不同而不同。这种形式的解释变量反映了*选项的特征*（*characteristics of the choices*），它们与通常的解释变量有所不同，后者不会随着选项的变化而变化。通常的解释变量反映的是*选择者的特征*（*characteristics of the chooser*），比如收入、教育程度，以及其他人口学特征。

McFadden（1974）提出了一个关于解释变量为选项特征的离散选择模型。他的模型也可以允许不同对象具有不同的可选集。例如，某些对象可能无法把地铁作为上班的交通工具。对于第 i 个对象以及结果变量的第 j 个选项，令 $\mathbf{x}_{ij} = (x_{ij1}, \cdots, x_{ijp})'$ 表示 p 个解释变量的取值，并令 $\mathbf{x}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{ip})'$。以第 i 个对象的可选集 C_i 为条件，关于选取第 j 个选项的概率的模型为

$$\pi_j(\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{ij})}{\sum_{h \in C_i} \exp(\boldsymbol{\beta}'\mathbf{x}_{ih})}。 \tag{7.22}$$

对于每两个选项 a 和 b，这个模型具有以下的 Logit 形式：

$$\log[\pi_a(\mathbf{x}_i)/\pi_b(\mathbf{x}_i)] = \beta'(\mathbf{x}_{ia} - \mathbf{x}_{ib})。 \tag{7.23}$$

以选择结果是 a 或 b 为条件，变量的效应取决于该对象在两个选项下这个变量的取值之间的距离。如果两个取值相等，那么模型认定这个变量对是否选择 a 或 b 没有影响。考虑到这种特性，McFadden 最初将式 7.22 模型称为*条件 Logit*（*conditonal Logit*）模型。

由式 7.23 可知，选择 a 而不是 b 的发生比不取决于可选集中的其他选项以及它们所对应的解释变量的取值。Luce（1959）将这种特性称为*独立于无关选项*（*independence from irrelevant alternatives*）。在某些应用中，这是不现实的。例如，关于交通工具的两个选项，自驾车和红色公车，假定 80% 选择自驾车，相应的发生比为 4.0。现在假设可选项为自驾车、红色公车和蓝色公车。按照式 7.23，选取自驾车而不是红色公车的发生比仍然是 4.0，但直觉上我们会预期它应当大约等于 8.0（选择每种公车的各占 10%）。McFadden（1974）写道："模型的适用范围应当局限于以下情况：在每个决策者的眼中，这些选项必须是截然不同的，并且决策者会独立权衡每个选项。"

7.6.2 离散选择与多项 Logit 模型

式 7.22 模型中也可以加入表示选择者特征的解释变量。这可能会令人惊讶,因为式 7.22 对每个解释变量都只有一个参数,也即,每两个选项所对应的参数向量是相同的。但是,式 7.2 多项 Logit 模型在将上述的解释变量替代为 J 个人为变量(artificial variables)后具有式 7.22 的离散选择形式,其中第 j 个变量是解释变量与一个当选择结果为 j 时等于 1 的虚拟变量的乘积。例如,对于单个的解释变量,令 x_i 表示对象 i 对应的取值。对于 $j = 1,\cdots,J$,当 $k = j$ 时令 δ_{jk} 等于 1,否则为 0,记

$$\mathbf{z}_{ij} = (\delta_{j1},\cdots,\delta_{jJ},\delta_{j1}x_i,\cdots,\delta_{jJ}x_i)'。$$

令 $\boldsymbol{\beta} = (\alpha_1,\cdots,\alpha_J,\beta_1,\cdots,\beta_J)'$,那么,$\boldsymbol{\beta}'\mathbf{z}_{ij} = \alpha_j + \beta_j x_i$,并且式 7.2 可表示为(为了模型的可识别性,令 $\alpha_J = \beta_J = 0$):

$$\pi_j(x_i) = \frac{\exp(\alpha_j + \beta_j x_i)}{\exp(\alpha_1 + \beta_1 x_i) + \cdots + \exp(\alpha_J + \beta_J x_i)}$$

$$= \frac{\exp(\boldsymbol{\beta}'\mathbf{z}_{ij})}{\exp(\boldsymbol{\beta}'\mathbf{z}_{i1}) + \cdots + \exp(\boldsymbol{\beta}'\mathbf{z}_{iJ})}。$$

它具有与式 7.22 模型相同的形式。

通过这种方法,离散选择模型可以既包括选择者的特征,也包括选项的特征。因此,式 7.22 模型是一个非常一般化的模型。使用基线类别 Logit 的普通多项 Logit 模型(式 7.2)是它的一个特例。

7.6.3 例子:购物地选择

McFadden(1974)利用多项 Logit 模型分析了宾夕法尼亚州匹兹堡的居民如何选择购物地。五种可能的选项对应于五个不同的市区。其中,一个解释变量测量的是购物机会,定义为该区的零售业就业占整个地区零售业就业的百分比。另一个解释变量是旅行成本,定义来源于对驾车所用时间和花费的一个单独分析。

以下是模型参数的最大似然估计结果:旅行成本为 -1.06 (SE =0.28),购物机会为 0.84 (SE =0.23)。由式 7.23 可得,

$$\log(\hat{\pi}_a/\hat{\pi}_b) = -1.06(P_a - P_b) + 0.84(S_a - S_b),$$

其中 P = 旅行成本, S = 购物机会。结果并不奇怪,随着旅行成本的下降以及购物机会的上升,购物地会变得相对更有吸引力。给定每个目的地对应的 P 和 S 的取值,式 7.22 给出了所估计的选取每个购物地的样本概率。

注 解

第 7.1 节:定类结果变量:基线类别 Logit 模型

7.1 多类别模型是基于对二分结果变量的潜变量方法的扩展演化而来的。一种方法使用了选取使效用最大化的类别的原则(习题 6.29)。Fahrmeir 和 Tutz(2001,Chap. 3)对此进行了讨论并给出了有关文献。对基线类别 Logit 模型的发展做出贡献的有 Bock(1970)、Haberman(1974a, pp. 352-373)、Mantel(1966)、Nerlove and Press (1973)、Theil(1969,1970)。Lesaffre 和 Albert(1989)介绍了有关的回归诊断。Amemiya(1981)、Haberman(1982),以及 Theil(1970)给出了相应的 R^2 指标。

第 7.2 节:定序结果变量:累积 Logit 模型

7.2 关于累积 Logit 模型的早期应用包括:Bock and Jones(1968)、Simon(1974)、Snell

(1964)、Walker and Duncan(1967)、Williams and Grizzle(1972)。McCullagh(1980)将比例发生比模型进行了推广。后期的有关文献包括:Agresti and Lang(1993a)、Hastie and Tibshirani(1987)、Peterson and Harrell(1990)、Tutz(1989)。另见第11.3.3节、注解 11.3,以及第 12.4.1 节。McCullagh 和 Nelder(1989,Sec. 5.6)建议使用累积总计(cumulative totals)来定义残差。

7.3 McCullagh(1980)指出,式 7.5 模型的计分检验等价于使用平均秩进行的非参数检验。例如,对于一个 $2 \times J$ 表格,假定 $\text{Logit}[P(Y \leqslant j)] = \alpha_j + \beta x$,其中 x 是一个指示变量。关于 $H_0: \beta = 0$ 的计分检验等价于一个离散情形下的 Wilcoxon-Mann-Whitney 检验。Whitehead(1993)给出了这种情况下的样本规模计算公式。为达到某个给定的统计效能水平,所需要的样本规模 n_J 随着 J 的增加而下降:当结果变量的类别具有相等的概率时,$n_J \approx 0.75n_2/(1-1/J^2)$。因此,在 J 较大的情况下,$n_J \approx 0.75n_2$,并且 $1-1/J^2$ 是一种使用 J 个类别而不是连续的结果变量的相对效率的测量指标。当 $J \approx 5$ 时,效率的损失是有限的;但是通过合并到 $J=2$ 时,会导致很大的效率损失。Edwardes(1997)通过将切点视作随机变量而创造性地应用了这个检验。这与第 12.4.1 节将要介绍的随机效应模型有关。

第 7.3 节:定序结果变量:累积连结模型

7.4 Aitchison 和 Silvey(1957)以及 Bock 和 Jones(1968,Chap. 8)讨论了累积 probit 模型。Farewell(1982)对补余双对数模型进行了扩展,使得模型允许在隐含刻度上的类别边界可以在样本间发生变动,这与随机效应模型有关(第 12.4 节)。Genter 和 Farewell(1985)介绍了一个一般性连结函数,它可以比较由 probit、补余双对数,以及其他连结函数所生成的模型的拟合情况。Yee 和 Wild(1996)界定了关于定类和定序结果变量的广义可加模型(generalized additive models)。Hamada 和 Wu(1990)以及 Nair(1987)介绍了检查式 7.8 模型的离散效应的其他方法。

7.5 一些研究者讨论了与随机排序有关的更一般性的统计推断,例如,参见:Dardanoni and Forcina(1998),以及有关的一些综述文章见:2002 issue of *J. Statist. Plann. Inference*(Vol. 107, Nos. 1-2)。

第 7.4 节:关于定序结果变量的其他模型

7.6 一个概率密度函数与它的累积分布函数的补函数之比被称为*历险函数*(*hazard function*)(第 9.7.3 节)。对于离散变量,它就是连续比 Logit 中的比。因此,连续比 Logit 有时被解释为历险的对数(log hazards)。Thompson(1977)在对离散生存数据进行模型分析时应用了这种方法。当时间的区间长度趋近于 0 时,它的模型转化为 Cox 比例风险模型。其他有关连续比 Logit 的应用包括:Läärä and Matthews(1985)、Tutz(1991)。

第 7.5 节:$I \times J \times K$ 表格中的条件独立性检验

7.7 令 $\mathbf{B}_k = \mathbf{u}_k \otimes \mathbf{v}_k$ 表示由 k 层中的行赋值 $\mathbf{u}_k$ 和列赋值 $\mathbf{v}_k$ 所形成的常数矩阵,其中 $\otimes$ 表示 Kronecker 乘积。Landis 等(1978)的广义统计量为

$$L^2 = \left[\sum_k \mathbf{B}_k(\mathbf{n}_k - \boldsymbol{\mu}_k)\right]' \left[\sum_k \mathbf{B}_k \mathbf{V}_k \mathbf{B}_k'\right]^{-1} \left[\sum_k \mathbf{B}_k(\mathbf{n}_k - \boldsymbol{\mu}_k)\right]。$$

当对所有的层都存在 $\mathbf{u}_k = (u_1, \cdots, u_I)$ 和 $\mathbf{v}_k = (v_1, \cdots, v_J)$ 时,$\text{L}^2 = \text{M}^2$。当 $\mathbf{u}_k$ 是 $(I-1) \times I$ 的矩阵$(\mathbf{I}, -1)$,其中 $\mathbf{I}$ 表示 $(I-1)$ 的单位矩阵,1 表示 $(I-1)$ 阶的元素为 1 的列向量,并且 $\mathbf{v}_k$ 也是相应的 $(J-1) \times J$ 矩阵时,L^2 简化为式 7.20,自由度为 $\text{df} = (I-1)(J-1)$。当 $\mathbf{u}_k$ 不变而 $\mathbf{v}_k = (v_1, \cdots, v_J)$ 时,L^2 是反映对所有层中 I 个

行均值与它们的零分布期望值相比较的总体指标,它的自由度为 df = (I − 1)。定序分类结果变量的秩赋值检验与对层进行调整后的 Spearman 相关系数和 Kruskal-Wallis检验相似。Landis 等(1998)以及 Stokes 等(2000)回顾了 CMH 方法。Koch 等(1982)也对有关方法进行了综述。

第 7.6 节:离散选择多项 Logit 模型

7.8 McFadden 的模型与 Bradley 和 Terry(1952)(参见第 10.6 节)以及 Luce(1959)所提出的模型有关。对此的文字描述,参见:Train(1986)。McFadden(1982)讨论了在树状结构中存在相互嵌套的选择时的多层次模型。有关的其他研究,参见:Maddala(1983)、Small(1987)。如果不假定独立于无关选择(independence from irrelevant alternatives),便会推导出一个 probit 连结的模型(Amemiya,1981),或者具有随机效应的 Logit 连结的模型(Brownstone and Train,1999)。对以上模型,可以利用第 12.6 节将要介绍的有关随机效应模型的方法拟合。这些方法包括决定似然函数的近似积分的蒙特卡洛法。有关综述,参见:Stern(1997)。

习 题

应用部分

7.1 对于表 7.13,令 Y = 信仰死后仍有来世,x_1 = 性别(1 = 女性,0 = 男性),x_2 = 种族(1 = 白人,0 = 黑人)。表 7.14 给出了对以下模型的拟合结果:

$$\log(\pi_j/\pi_3) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, \quad j = 1,2,$$

括号中数值为相应的标准误。

表 7.13 习题 7.1 的数据

种 族	性 别	信仰死后仍有来世		
		是	不确定	否
白人	女性	371	49	74
	男性	250	45	71
黑人	女性	64	9	15
	男性	25	5	13

来源:1991 年美国综合社会调查,美国民意研究中心(National Opinion Research Center)。

表 7.14 习题 7.1 的模型拟合结果

参 数	有关信仰类别的 Logit	
	是/否	不确定/否
截距	0.883(0.243)	−0.758(0.361)
性别	0.419(0.171)	0.105(0.246)
种族	0.342(0.237)	0.271(0.354)

a. 给出关于 $\log(\pi_1/\pi_2)$ 的预测方程。

b. 根据回答"是"和"否"的类别,通过有关发生比之比的 95% 的置信区间,解释性别的条件效应。

c. 对于白人女性,给出 $\hat{\pi}_1 = \hat{P}(Y = 是) = 0.76$ 。

d. 不计算估计概率,说明为什么截距的估计值表明对于黑人男性存在 $\hat{\pi}_1 > \hat{\pi}_3 > \hat{\pi}_2$ 。通过截距和性别效应的估计值,证明上述排序对黑人女性也适用。

e. 不计算估计概率，说明为什么性别和种族效应的估计值表明黑人男性的 $\hat{\pi}_3$ 最大。

f. 关于这个拟合，$G^2 = 0.9$。说明为什么残差自由度为 df = 2。去除掉性别效应后，$G^2 = 8.0$。检验在给定种族后，信仰的选择是否独立于性别。对结果加以解释。

7.2 一个模型利用 x = 年收入（以 10 000 美元为单位）来预测对美国总统的偏好（民主党 D，共和党 R，独立党派 I）。它的拟合结果为 $\log(\hat{\pi}_D/\hat{\pi}_I) = 3.3 - 0.2x$ 以及 $\log(\hat{\pi}_R/\hat{\pi}_I) = 1.0 + 0.3x$。

a. 给出关于 $\log(\hat{\pi}_R/\hat{\pi}_D)$ 的预测方程，并对斜率加以解释。当 x 取哪些值时，$\hat{\pi}_R > \hat{\pi}_D$？

b. 给出关于 $\hat{\pi}_I$ 的预测方程。

c. 当 x 的取值范围为 0 到 10 之间时，分别画出 $\hat{\pi}_D$，$\hat{\pi}_I$ 以及 $\hat{\pi}_R$ 的曲线，并加以解释。

7.3 表 7.15 所示为性别和种族对政党认同的效应。找出一个能够很好拟合该数据的基线类别 Logit 模型。解释所估计的对于政党认同是民主党而不是共和党的发生比的效应。

表 7.15 习题 7.3 的数据

性别	种族	政党认同		
		民主党	共和党	独立党派
男性	白人	132	176	127
	黑人	42	6	12
女性	白人	172	129	130
	黑人	56	4	15

7.4 对于在佛罗里达乔治湖中捕获的 63 只短吻鳄，表 7.16 给出了它们的主要觅食类型（鱼类，无脊椎动物，其他）和长度（米）。如果短吻鳄的长度 < 1.83 米（6 英尺）就被视为幼鳄，长度 > 1.83 米则视为成年鳄。

表 7.16 习题 7.4 的数据[a]

雄性				雌性			
长度/m	觅食种类	长度/m	觅食种类	长度/m	觅食种类	长度/m	觅食种类
1.30	*I*	1.80	*F*	1.24	*I*	2.56	*O*
1.32	*F*	1.85	*F*	1.30	*I*	2.67	*F*
1.32	*F*	1.93	*I*	1.45	*I*	2.72	*I*
1.40	*F*	1.93	*F*	1.45	*O*	2.79	*F*
1.42	*I*	1.98	*I*	1.55	*I*	2.84	*F*
1.42	*F*	2.03	*F*	1.60	*I*		
1.47	*I*	2.03	*F*	1.60	*I*		
1.47	*F*	2.31	*F*	1.65	*F*		
1.50	*I*	2.36	*F*	1.78	*I*		
1.52	*I*	2.46	*F*	1.78	*O*		
1.63	*I*	3.25	*O*	1.80	*I*		
1.65	*O*	3.28	*O*	1.88	*I*		
1.65	*O*	3.33	*F*	2.16	*F*		
1.65	*I*	3.56	*F*	2.26	*F*		
1.65	*F*	3.58	*F*	2.26	*F*		
1.68	*F*	3.66	*F*	2.36	*F*		

续表

雄性				雌性			
长度/m	觅食种类	长度/m	觅食种类	长度/m	觅食种类	长度/m	觅食种类
1.70	*I*	3.68	*O*	2.39	*F*		
1.73	*O*	3.71	*F*	2.41	*F*		
1.78	*F*	3.89	*F*	2.44	*F*		
1.78	*O*						

a *I*,无脊椎动物;*F*,鱼类;*O*,其他。

a. 按照(成年鳄,幼鳄)来表示长度,建立一个能够充分描述性别和长度对觅食类型的效应的模型。对这些效应加以解释。对于成年雌性短吻鳄,求出选择各个食物类别的估计概率。

b. 仅利用觅食选择是鱼类或无脊椎动物的观测值,建立一个能够充分描述性别和长度(二分变量)的效应的模型。将这种通过单独拟合获得的参数估计和标准误与包括“其他”类别在内同时拟合的结果进行比较。

c. 将长度视为二分变量会损失一些信息。将长度作为连续变量重新拟合(a)部分的模型,解释并说明所估计的结果概率如何随长度的变动而变动。求当“无脊椎动物”和“其他”发生的可能性相等时所估计的长度。

7.5 在一次近期的综合社会调查(General Social Survey)数据中,以 Y = 政治理想(非常开放,比较开放,中等,比较保守,非常保守),以及 $x = 1$ 代表428名民主党员,$x = 0$ 代表407名共和党员。对该数据拟合式7.5累积Logit模型的结果有 $\hat{\beta} = 0.975$ (SE = 0.129),$\hat{\alpha}_1 = -2.469$。解释 $\hat{\beta}$。求每个党派中回答结果为“非常开放”的估计概率。

7.6 参见习题7.5。拟合相邻类别Logit模型,得出 $\hat{\beta} = 0.435$。利用相邻类别的发生比之比来解释“非常开放”和“非常保守”这两种类别的关系。

7.7 表7.17是关于第8.4.2节所分析的数据的一个扩展版。结果变量的类别为:①未受伤,②受伤了但没动用救护车,③受伤了并且动用了救护车但没住院,④受伤并住院但没死亡,⑤受伤并死亡。表7.18给出了一个类似于式7.5的模型的拟合结果,其中预测变量由虚拟变量表示。

a. 为什么这里有四个截距项?以城区系安全带的男性为例,说明这些截距项如何决定了结果变量的估计分布。

b. 在给定使用安全带和地理位置的情况下,构建关于性别效应的置信区间。加以解释。

c. 给定性别,对于农村地区和城市地区,分别估计使用安全带与结果变量之间的累积发生比之比。基于以上结果,说明使用安全带的效应如何在不同地区间变动,并说明如何解释有关交叉项的估计结果(-0.124 4)。

表7.17 习题7.7的数据

性别	地区	是否使用安全带	结果				
			1	2	3	4	5
女性	城市	否	7 287	175	720	91	10
		是	11 587	126	577	48	8
	农村	否	3 246	73	710	159	31
		是	6 134	94	564	82	17

续表

性别	地区	是否使用安全带	结果				
			1	2	3	4	5
男性	城市	否	10 381	136	566	96	14
		是	10 969	83	259	37	1
	农村	否	6 123	141	710	188	45
		是	6 693	74	353	74	12

来源：数据由缅因州奥古斯塔(Augusta, Maine)医疗发展中心的 Cristanna Cook 友情提供。

表 7.18　习题 7.7 的模型结果

参数			DF	估计值	标准误
截距 1			1	3.307 4	0.035 1
截距 2			1	3.481 8	0.035 5
截距 3			1	5.349 4	0.047 0
截距 4			1	7.256 3	0.091 4
性别	女性		1	-0.546 3	0.027 2
性别	男性		0	0.000 0	0.000 0
地区	农村		1	-0.698 8	0.042 4
地区	城市		0	0.000 0	0.000 0
安全带	否		1	-0.760 2	0.039 3
安全带	是		0	0.000 0	0.000 0
地区 * 安全带	农村	否	1	-0.124 4	0.054 8
地区 * 安全带	农村	是	0	0.000 0	0.000 0
地区 * 安全带	城市	否	0	0.000 0	0.000 0
地区 * 安全带	城市	是	0	0.000 0	0.000 0

7.8　参见表 7.8 的累积 Logit 模型。

a. 比较估计的收入效应 $\hat{\beta}_1 = -0.510$ 与通过合并类别(i)非常满意和比较满意以及(ii)非常不满意和有点不满意之后对三个类别的估计结果。这反映了该模型的什么特性?

b. 考虑使用结果变量所有类别时的 $\hat{\beta}_1$/SE 以及在(a(i))部分合并后的情况。一般来说,合并结果变量的多项类别的缺点之一是会导致效应显著性的下降。

c. 检查允许收入和性别之间的交互效应是否会改进模型的拟合,加以解释。

7.9　表 7.19 所示为一项关于治疗小细胞肺癌的临床试验。病人被随机分配到两个干预组。顺序疗法在每个干预周期内都使用了同样配方的化学药剂;交替疗法使用了三种不同配方的化学药剂,每个周期都有变动。

表 7.19　习题 7.9 的数据

疗法	性别	化疗效果			
		继续恶化	无变化	局部好转	全面好转
顺序	男性	28	45	29	26
	女性	4	12	5	2
交替	男性	41	44	20	20
	女性	12	7	3	1

来源:W. Holtbrugge and M. Schumacher, *Appl. Statist.* 40:249-259(1991)。

a. 拟合一个包括干预方式和性别的主效应的累积 Logit 模型，加以解释。

b. 拟合一个还包括交互项的模型，加以解释。它的拟合结果更好吗？说明为什么它等价于利用性别-干预方式的四个取值组合作为单个变量的值的情况。

7.10 参考表 7.13。将相信死后有来生当作定序变量，拟合一个定序模型，并加以解释。

7.11 表 9.7 给出了一些工厂工人的吸烟状况（S）、呼吸检查结果（B），以及年龄（A）之间的关联。将 B 作为结果变量。

a. 构建一个基线类别 Logit 模型，将 S 和 A 当作具有可加效应的分类变量。这个模型的偏离度为 $G^2 = 25.9$。说明它的自由度 df = 4，并解释为什么这个模型将所有变量都视为定类变量。

b. 将 B 作为定序变量，将 S 也视作定序变量以测度最近吸烟是什么时候，并赋值为 $\{s_i\}$。考虑以下模型：

$$\log \frac{P(B = k+1 \mid S = i, A = j)}{P(B = k \mid S = i, A = j)} = \alpha_k + \beta_1 s_i + \beta_2 a_j + \beta_3 s_i a_j,$$

其中 $a_1 = 0, a_2 = 1$。证明这个模型假定 S 具有线性效应，当年龄 < 40 时，它的斜率为 β_1；当年龄在 40 ~ 59 岁时，它的斜率为 $\beta_1 + \beta_3$。利用 $\{s_i = i\}$，得出 $\hat{\beta}_1 = 0.115$，$\hat{\beta}_2 = 0.311$，以及 $\hat{\beta}_3 = 0.663$（SE = 0.164）。解释这个交互效应。

c. 根据(b)部分，证明对于目前吸烟者检查结果为不正常而不是基本正常的估计发生比是曾吸烟者的 2.18 倍，是从未吸烟者的 $\exp(2 \times 0.778) = 4.74$ 倍。说明为什么这些值的平方是所估计的关于呼吸检查结果为不正常而不是正常的发生比。

7.12 本书的网站（*www.stat.ufl.edu/~aa/cda/cda.html*）包括一个关于 1965 年高中毕业生的 7 × 2 表格。数据按照他们是否至少参加过一次游行、抗议或静坐被划分为是否为抗议者，同时数据也给出了他们在 1982 年时的政党认同。使用(a)政党认同、(b)是否是一个抗议者作为结果变量，对数据进行分析。解释结果，并加以比较。

7.13 对于表 7.5，累积 probit 模型的拟合结果为 $\Phi^{-1}[\hat{P}(Y \leqslant j)] = \hat{\alpha}_j - 0.195x_1 + 0.683x_2$，其中 $\hat{\alpha}_1 = -0.161, \hat{\alpha}_2 = 0.746, \hat{\alpha}_3 = 1.339$。在 x_2 = 社会经济地位的两个取值水平上给出，作为 x_1 = 生命事件指数的函数的曲线 $\hat{P}(Y > 2)$ 的两个累积分布函数的均值和标准差。解释有关的效应。

7.14 利用累积 probit 模型来分析表 7.8。解释结果，并与书中其他定序模型的结果加以比较。

7.15 用表 7.20 中美国东北部家庭的家庭收入分布百分比数据，拟合一个补余双对数模型。解释收入分布之间的差异。

表 7.20　习题 7.15 的数据

年　份	收　入(1 000 美元)						
	0~3	3~5	5~7	7~10	10~12	12~15	15+
1960	6.5	8.2	11.3	23.5	15.6	12.7	22.2
1970	4.3	6.0	7.7	13.2	10.5	16.3	42.1

来源：授权重印自：the Royal Statistical Society, London（McCullagh 1980）。

7.16 表 7.21 是对表 7.8 的数据拟合平均结果变量模型的结果，其中关于收入的赋值为 $\{3, 10, 20, 35\}$，关于工作满意度的赋值为 $\{1, 3, 4, 5\}$。解释收入的效应，在控制

性别的情况下，给出在收入水平分别为35和3时的平均满意度之差的置信区间，并检查模型的拟合。

表7.21 习题7.16的结果

Source	DF	Chi-Square	Pr > ChiSq
Residual	5	6.99	0.221 1

Analysis of Weighted Least Squares Estimates

Effect	Parameter	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	1	3.807 6	0.179 6	449.47	<.000 1
gender	2	-0.068 7	0.141 9	0.23	0.628 3
income	3	0.016 0	0.006 6	5.97	0.014 6

译者注——Gender：性别；Income：收入。

7.17 本书的网站(*www. stat. ufl. edu/ ~aa/cda/cda. html*)包括一个3×4×4的表格，它对四所医院(H)的十二指肠溃疡患者在手术(X)后的腹泻严重程度(Y)进行了交叉分组。四种手术指的是对病人的不同治疗方式，它们具有自然的排序。腹泻严重程度描述的是手术所可能导致的副作用，它的三个类别也是排序的。

a. 表7.22给出了广义CMH检验的结果。解释并说明为什么其中一种检验的显著性水平会比其他检验高得多。

表7.22 习题7.17的结果

Summary Statistics for dumping by operate

Controlling for hospital

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.340 4	0.011 8
2	Row Mean Scores Differ	3	6.590 1	0.086 2
3	General Association	6	10.598 3	0.101 6

b. 令$\{x_i = i\}$。拟合模型

$$\text{Logit}[P(Y \leqslant j \mid H = h, X = i)] = \alpha_j + \mu_h + \beta x_i。$$

利用该模型检验X和Y的条件独立性，并解释$\hat{\beta}$。哪一个广义CMH检验具有与此相同的效果？

c. 通过允许手术的效应在不同医院之间变动能提高模型的拟合程度吗？加以解释。

d. 构建一个能够很好拟合该数据的平均结果变量模型。加以解释。

7.18 表7.23的研究中将对象随机分配到控制组或干预组。在研究期间，干预组的对象每天食用包含欧车前的谷类。该研究分析饮食对LDL胆固醇水平的效应。

表7.23 习题7.18的数据

开始时的水平	结束时的LDL胆固醇水平							
	控制组				干预组			
	≤3.4	3.4~4.1	4.1~4.9	>4.9	3.4	3.4~4.1	4.1~4.9	>4.9
≤3.4	18	8	0	0	21	4	2	0
3.4~4.1	16	30	13	2	17	25	6	0
4.1~4.9	0	14	28	7	11	35	36	6
>4.9	0	2	15	22	1	5	14	12

来源：数据由美国家乐氏公司(Kellogg Co.)的Sallee Anderson友情提供。

a. 利用开始时的LDL胆固醇水平作为协变量，将结束时的胆固醇水平作为干预方式的函数进行模型分析。解释干预的效应。

b. 现在将开始时的水平作为分类变量,重做(a)部分。对结果加以比较。

c. 关于(b)部分的另一种方法是,在根据开始时胆固醇水平划分的分表中,对有关干预方式与结束时水平进行广义 CMH 检验。考虑结果变量的排序特征,利用该检验考察干预方式的效应。解释结果,并与(b)部分加以比较。

7.19 利用本章所介绍的各种模型分析表 7.5。撰写一份报告,综述分析结果,并比较各种模型之间的优缺点。

7.20 本书的网站(*www.stat.ufl.edu/~aa/cda/cda.html*)包括一个 4×4×5 的表格,它对认知损伤评估、老年痴呆症,以及年龄进行了交叉分组。分别将(a)老年痴呆症和(b)认知损伤作为结果变量,对这些数据进行分析。

7.21 利用 Logit 模型分析表 9.5 的数据,分别将(a)政党认同和(b)意识形态作为结果变量。

7.22 本书的网站(*www.stat.ufl.edu/~aa/cda/cda.html*)包括一个关于哥本哈根居民的样本的 4×2×3×3 的表格。它的变量分别是居住方式(H)、与其他住户的联系(C)、感觉对公寓管理的影响力(I),以及对居住条件的满意度(S)。将 S 作为结果变量,对这些数据进行分析。

7.23 参考表 7.17。分析这些数据。

理论与方法

7.24 一个关于指数离散分布族(exponential dispersion family)(式 4.14)的多元扩展为

$$f(\mathbf{y}_i;\boldsymbol{\theta}_i,\phi) = \exp\{[\mathbf{y}_i'\boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)]/a(\phi) + c_i(\mathbf{y}_i,\phi)\},$$

其中 $\boldsymbol{\theta}_i$ 是自然参数(natural parameter)。证明在第 7.1.5 节定义的参数为 $\{\pi_j, j = 1,\cdots,J-1\}$ 的一次单一试验的多项分布变量 $\mathbf{y}_i$ 属于具有 $(J-1)$ 个参数的指数分布族,其基线类别 Logit 就是它的自然参数。

7.25 在一个 $I\times J$ 列联表中,单元格计数 $\{y_{ij}\}$ 服从参数为 $(n;\{\pi_{ij}\})$ 的多项分布。证明 $\{P(Y_{ij} = n_{ij}), i = 1,\cdots,I, j = 1,\cdots,J\}$ 可以被表示为

$$d^n n!\prod_i\prod_j(n_{ij}!)^{-1}\exp\Bigg[\sum_{i=1}^{I-1}\sum_{j=1}^{J-1}n_{ij}\log(\alpha_{ij}) + \sum_{i=1}^{I-1}n_{i+} + \log(\pi_{iJ}/\pi_{IJ})\Bigg] + \sum_{j=1}^{J-1}n_{+j}\log(\pi_{Ij}/\pi_{IJ}),$$

其中 $\alpha_{ij} = \pi_{ij}\pi_{IJ}/\pi_{iJ}\pi_{Ij}$,并且 d 是一个独立于数据的常数。通过证明以下方程,

$$\sum_i\sum_j n_{ij}\log\alpha_{ij} = \sum_i\sum_j s_{ij}\log\theta_{ij},\quad 其中\quad s_{ij} = \sum_{a\leq i}\sum_{b\leq j}n_{ab},$$

将其表示为局部发生比之比 $\{\theta_{ij}\}$ 的函数。

7.26 假定我们将式 7.2 表示为

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x})}{\sum_{h=1}^{J}\exp(\alpha_h + \boldsymbol{\beta}_h'\mathbf{x})}。$$

证明对分子和分母分别除以 $\exp(\alpha_J + \boldsymbol{\beta}_J'\mathbf{x})$ 得出的新参数为 $\alpha_j^* = \alpha_j - \alpha_J$ 和 $\boldsymbol{\beta}_j^* = \boldsymbol{\beta}_j - \boldsymbol{\beta}_J$,并且它们满足 $\alpha_J = 0$ 和 $\boldsymbol{\beta}_J = \mathbf{0}$。因此,限定 $\alpha_J = 0$ 和 $\boldsymbol{\beta}_J = \mathbf{0}$ 不会影响模型的一般性。

7.27 当 $J = 3$ 时,假定

$$\pi_j(x) = \exp(\alpha_j + \beta_j x)/[1 + \exp(\alpha_1 + \beta_1 x) + \exp(\alpha_2 + \beta_2 x)],$$

$j = 1,2$。证明 $\pi_3(x)$ 满足:(a)如果 $\beta_1 > 0$ 且 $\beta_2 > 0$,它是 x 的减函数,(b)如果 $\beta_1 < 0$ 且 $\beta_2 < 0$,它是 x 的增函数,(c)如果 β_1 和 β_2 方向相反,它不单调。

7.28 参考有关基线类别 Logit 模型的对数似然函数(第 7.1.4 节)。将充分统计量表述为 $np_j = \sum_i y_{ij}$ 和 $S_{jk} = \sum_i x_{ik} y_{ij}, j = 1, \cdots, J-1, k = 1, \cdots, p$。令 $\mathbf{S} = (S_{11}, \cdots, S_{1t}, \cdots, S_{J1}, \cdots, S_{Jt})'$。以 $\sum_i y_{ij}$ 为条件, $j = 1, \cdots, J$, 在解释变量不存在效应的零假设下,证明

$$E(\mathbf{S}) = n(\mathbf{P} \otimes \mathbf{m}), \quad \text{var}(\mathbf{S}) = n(\mathbf{V} \otimes \sum),$$

其中 $\mathbf{p} = (p_1, \cdots, p_J)'$; $\mathbf{m} = (\bar{x}_1, \cdots, \bar{x}_t)'$, 这里 $\bar{x}_k = (\sum_i x_{ik})/n$; $\sum$ 的元素为 (s_{kv}^2), 这里 $s_{kv}^2 = [\sum_i (x_{ik} - \bar{x}_k)(x_{iv} - \bar{x}_v)]/(n-1)$; $\mathbf{V}$ 的元素为 $v_{ii} = p_i(1-p_i)$ 和 $v_{ij} = -p_i p_j$, 并且 $\otimes$ 表示 Kronecker 乘积(Zelen, 1991)。

7.29 比例发生比模型是基线类别 Logit 模型的一个特例吗? 说明原因。

7.30 给出关于多项分布的因式分解(式 7.15)的证明。

7.31 关于模型 $\text{Logit}[P(Y \leq j)] = \alpha_j + \beta_j x$, 证明对某些 x 值,累积概率可能出现排序混乱的情形。

7.32 在一个 $I \times J$ 列联表中, Y 为定序变量, x 的赋值为 $\{x_i = i\}$, 考虑模型

$$\text{Logit}[P(Y \leq j \mid X = x_i)] = \alpha_j + \beta x_i \text{。} \quad (7.24)$$

a. 证明 $\text{Logit}[P(Y \leq j \mid X = x_{i+1})] - \text{Logit}[P(Y \leq j \mid X = x_i)] = \beta$。证明这个 Logit 之差是一个 2×2 表格的对数累积发生比之比,该表格由行 i 和行 $i+1$ 以及以 j 作为切点的二分结果变量组成。因此,式 7.24 是关于累积发生比之比的一致性关联模型(*uniform association model*)。

b. 证明残差的自由度为 $\text{df} = IJ - I - J$。

c. 证明 X 和 Y 相互独立对应于 $\beta = 0$ 时的特例。

d. 保持线性预测项不变,利用相邻类别 Logit 模型,证明局部发生比之比(式 2.10)存在一致性关联。

e. 对式 7.24 的一个扩展将 $\{\beta x_i\}$ 替代为不具有排序特征的参数 $\{\mu_i\}$, 因而将 X 视为定类变量。在行 a 和 b 中,证明对于所有 $J-1$ 个切点,对数累积发生比之比都等于 $\mu_a - \mu_b$。

7.33 假定式 7.24 模型对于一个 $J > 2$ 的 $2 \times J$ 表格成立,并且令 $x_2 - x_1 = 1$。说明为什么局部对数发生比之比的绝对值一般要比累积对数发生比之比 β 小(事实上, McCullagh 和 Nelder(1989)指出,局部发生比之比 $\{\theta_{ij}\}$ 和 β 的关系为

$$\log\theta_{1j} = \beta[P(Y \leq j+1) - P(Y \leq j-1)] + o(\beta), \quad j = 1, \cdots, J-1,$$

其中当 $\beta \to 0$ 时, $o(\beta)/\beta \to 0$)。

7.34 某一结果变量包括以下类别(强烈同意,比较同意,比较不同意,非常不同意,不知道)。对这样的变量进行模型分析的一种方法是,使用 Logit 模型分析"不知道"发生的概率,而在结果变量取值落在其他类别的条件下通过定序模型来分析这些有序的类别。说明如何构建一个能同时拟合上述两个模型的似然函数。

7.35 关于累积 probit 模型 $\Phi^{-1}[P(Y \leq j)] = \alpha_j - \boldsymbol{\beta}'\mathbf{x}$, 在控制其他预测变量的情况下,说明为什么 x_i 每增加 1 单位对应着隐含的潜变量预计增加 β_i 个标准差。

7.36 在式 7.7 累积连结模型中,证明对于 $1 \leq j < k \leq J-1$, $P(Y \leq k \mid \mathbf{x}) = P(Y \leq j \mid \mathbf{x}^*)$, 其中 $\mathbf{x}^*$ 为将 $\mathbf{x}$ 的第 i 个元素加上 $(\alpha_k - \alpha_j)/\beta_i$。加以解释。

7.37 在包括一个分类预测变量的情况下,关于 $I \times J$ 列联表的累积连结模型为

$$G^{-1}[P(Y \leq j)] = \alpha_j + \mu_i, \quad i = 1, \cdots, I, \quad j = 1, \cdots, J-1\text{。}$$

a. 证明残差自由度为 $\text{df} = (I-1)(J-2)$。

b. 当这个模型成立时,证明独立性对应着 $\mu_1 = \cdots = \mu_I$, 并且独立性检验的自由度

为 df $= I-1$。

c. 当这个模型成立时，证明对 Y 来说，各行之间是随机排序的。

7.38 当 $y > 0$ 时，$F_1(y) = 1 - \exp(-\lambda y)$ 是一个参数为 λ 的负指数累积分布函数，并且 $F_2(y) = 1 - \exp(-\mu y)$。证明对于所有 y，补余双对数累积分布函数之间的差都相等。给出这一结果对于分类数据分析的含义。

7.39 考虑以下模型：连结$[\omega_j(\mathbf{x})] = \alpha_j + \boldsymbol{\beta}_j' x$，其中 $\omega_j(\mathbf{x})$ 为式 7.14。

a. 说明为什么这个模型可以对 $j = 1, \cdots, J-1$ 分别进行拟合。

b. 对于补余双对数连结，证明这个模型等价于对累积概率使用相同的连结（Läärä and Matthews，1985）。

7.40 对于定序结果变量，为什么如在正态分布回归中那样通过最小二乘法拟合平均结果变量模型不是最优的？

7.41 当 X 和 Y 都是定序变量时，解释如何通过允许在每个分表中具有不同的趋势来检验条件独立性（提示：利用 β_k 替代 β，对式 7.17 模型进行扩展）。

7.42 某一小餐馆提供四种主菜：鸡肉，牛肉，鱼肉，素食。利用 x = 性别（1 = 女性，0 = 男性）和 u = 主菜价格，构建一个类似式 7.22 的关于主菜选择的模型，其中 u 是选项的特征。对模型的参数加以解释。

8 关于列联表的对数线性模型

在第4.3节我们提到了对数线性模型,并指出它是对服从泊松分布的结果变量使用对数连结函数的一种广义线性模型。对数线性模型的一种常见应用是,对列联表中的单元格计数进行模型分析。模型设定单元格的期望计数取决于形成该单元格的分类变量的取值以及这些变量之间的关联和交互效应。对数线性模型的目的就是分析变量之间的关联和交互模式。

在第8.1节,我们介绍关于二维列联表的对数线性模型。在第8.2和8.3节,我们将其扩展到三维表格的情况。第8.4节讨论关于多维表格的模型。对数线性模型主要应用于至少有两个变量是结果变量的情况。如果只有一个分类结果变量,那么更简单、更自然的方法是应用Logit模型。当一个变量是结果变量而其他变量被作为解释变量时,关于该结果变量的Logit模型等价于某一特定的对数线性模型。第8.5节探讨Logit模型与对数线性模型之间的这种联系。在第8.6和8.7节,我们介绍关于对数线性模型的最大似然拟合。

8.1 关于二维表格的对数线性模型

考虑将一个多项分布样本的 n 个对象按照两个分类结果变量进行交叉划分所形成的 $I \times J$ 列联表。它的单元格概率为 $\{\pi_{ij}\}$,相应的期望频数等于 $\{\mu_{ij} = n\pi_{ij}\}$。对数线性模型的公式使用 $\{\mu_{ij}\}$ 而不是 $\{\pi_{ij}\}$,因而该模型也适用于 $N = IJ$ 个独立单元格计数 $\{Y_{ij}\}$ 的泊松分布样本,其中 $\{\mu_{ij} = E(Y_{ij})\}$。无论在哪种情况下,我们都将单元格的观察计数表示为 $\{n_{ij}\}$。

8.1.1 独立性模型

在第4.3.6节中我们指出,在统计独立性的情况下,$\{\mu_{ij}\}$ 具有以下结构:

$$\mu_{ij} = \mu\alpha_i\beta_j。$$

例如,对于多项分布样本,$\mu_{ij} = n\pi_{i+}\pi_{+j}$。以 X 表示行变量,Y 表示列变量。独立性的表达公式是可积的。因此,对于行效应 λ_i^X 和列效应 λ_j^Y,$\log\mu_{ij}$ 具有可加的形式:

$$\log\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y。\tag{8.1}$$

这个模型是*独立性对数线性模型*(*loglinear model of independence*)。与通常情况一样,模型的可识别性要求附加诸如 $\lambda_I^X = \lambda_J^Y = 0$ 的约束条件。

独立性模型的最大似然拟合值为 $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$,即卡方独立性检验所估计的期望

频数。相应的 X^2 和 G^2 独立性检验(第 3.2.1 节)也是这个对数线性模型的拟合优度检验。

8.1.2 参数的解释

关于列联表的对数线性模型属于广义线性模型,它的 N 个单元格计数被视为服从一个泊松随机分布的独立观测值。作为广义线性模型,对数线性模型视列联表数据为 N 个单元格计数,而不是对 n 个个体对象的交叉划分。单元格的期望计数与模型的解释变量之间的连结函数为对数函数。如式 8.1 所示,对于构成列联表的两个变量,模型并不区分结果变量和解释变量。它将两个变量的联合分布作为结果变量,通过模型分析它们的取值组合点上的 $\{\mu_{ij}\}$。但是,在对参数进行解释时,将其中一个变量视为结果变量会更方便一些。

我们以 $I\times 2$ 表格的独立性模型为例来具体说明。在第 i 行中,logit 等于

$$\begin{aligned}\text{logit}[P(Y=1\mid X=i)] &= \log\frac{P(Y=1\mid X=i)}{P(Y=2\mid X=i)}\\ &= \log\frac{\mu_{i1}}{\mu_{i2}} = \log\mu_{i1} - \log\mu_{i2}\\ &= (\lambda+\lambda_i^X+\lambda_1^Y)-(\lambda+\lambda_i^X+\lambda_2^Y) = \lambda_1^Y-\lambda_2^Y。\end{aligned}$$

上式最后一项不取决于 i,也即,$\text{logit}[P(Y=1\mid X=i)]$ 在 X 的每个取值上都相等。因此,独立性意味着模型具有形式 $\text{logit}[P(Y=1\mid X=i)]=\alpha$。在每一行中,结果落在第 1 列的发生比都等于 $\exp(\alpha)=\exp(\lambda_1^Y-\lambda_2^Y)$。

当 $J>2$ 时,也存在类似的特性。对于一个给定的变量,两个参数之差取决于对该变量选择一种结果而不是其他结果的对数发生比。当然,在只有一个结果变量的情况下,可以直接应用 Logit 模型,没有必要使用对数线性模型。

8.1.3 饱和模型

对于统计上相依的变量,可以利用一个更复杂的对数线性模型来分析,

$$\log\mu_{ij}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_{ij}^{XY}, \tag{8.2}$$

其中,$\{\lambda_{ij}^{XY}\}$ 作为关联项反映了对独立性的偏离。式 8.2 模型的等式右边与关于单元格均值的二维方差分析(two-way ANOVA)中允许交互效应的公式很相似。$\{\lambda_{ij}^{XY}\}$ 代表 X 和 Y 之间的交互效应,即一个变量对 μ_{ij} 的效应取决于另一个变量的取值。当存在所有 $\lambda_{ij}^{XY}=0$ 时,就得出式 8.1 独立性模型。

在式 8.1 和式 8.2 中以 $\lambda_I^X=\lambda_J^Y=0$ 为约束条件,$\{\lambda_i^X\}$ 和 $\{\lambda_j^Y\}$ 等价于代表 X 的前 $(I-1)$ 个类别和 Y 的前 $(J-1)$ 个类别的虚拟变量的系数。因此,λ_{ij}^{XY} 是虚拟变量 λ_i^X 和 λ_j^Y 的乘积项的系数。因为这里存在 $(I-1)(J-1)$ 个交叉乘积项,$\lambda_{Ij}^{XY}=\lambda_{iJ}^{XY}=0$,所以这些参数中只有 $(I-1)(J-1)$ 个是非冗余的。独立性检验就是分析是否这 $(I-1)(J-1)$ 个参数等于零,因而检验的残差自由度为 $\text{df}=(I-1)(J-1)$。

式 8.2 模型中的参数数量等于 $1+(I-1)+(J-1)+(I-1)(J-1)=IJ$,也即单元格的数量。因此,这个模型可以完全地描述任意的 $\{\mu_{ij}>0\}$(参见习题 8.16)。它是关于二维列联表的最一般性的模型,即*饱和模型*(*saturated model*)。对此模型,对数发生比之比与 $\{\lambda_{ij}^{XY}\}$ 之间存在直接的关系。例如,对于 2×2 表格,

$$\log\theta=\log\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}=\log\mu_{11}+\log\mu_{22}-\log\mu_{12}-\log\mu_{21}$$

$$= (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) -$$
$$(\lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY})$$
$$= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}。\quad (8.3)$$

因此,$\{\lambda_{ij}^{XY}\}$决定了相应的关联。

在实际应用中,我们更偏好非饱和模型,这是因为非饱和模型的拟合对样本数据进行了修匀,且解释起来更为简单。对于三维及以上的表格,非饱和模型中也可以包括关联项。这时,我们用对数线性模型来更多地描述关联(即二维项)而不是发生比(对应于单维项)。

与本书中的其他模型相同,式 8.2 模型是*分级的*(*hierarchical*)。这意味着,模型在包括一个高阶项时必须同时包括这些变量的所有低阶项。当模型包括 λ_{ij}^{XY}时,它也会包括 λ_i^X 和 λ_j^Y。包括这些低阶项的原因在于,如果不这样做的话,高阶项的统计显著性以及对它的解释将会取决于变量是如何编码的。我们不希望出现这种情况,而在分级的模型中,无论变量如何编码,结果都是一样的。

不分级模型(nonhierarchical model)的一个例子是

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_{ij}^{XY}。$$

这个模型允许存在关联项,但是它的期望频数具有一些奇怪的特性,即其模式取决于对参数所使用的约束条件。例如,当约束条件设定变量最后一个类别所对应的参数为零时,每列中都有 $\log \mu_{Ij} = \lambda$。在实际应用中,不分级的模型往往都不具有合理性。这些模型相当于在方差分析或回归分析中,包括交互项却不包括相应的主效应的情况。

当模型中包括二维项时,关于模型的解释会主要关注这些高阶项而不是单维项。与具有二维交互项的二维方差分析相似,这时仅解释主效应可能会得出误导性的结论。对主效应的估计取决于高阶项中所使用的编码规则,并且对它的解释也取决于编码本身(参见习题 8.16)。如在第 8.2 节所讨论的,一般来说,我们会将注意力限定在某个变量的高阶项上。

8.1.4　其他的参数约束条件

与独立性模型一样,饱和模型中的参数约束条件也具有随意性。比如,除了设定所有的 $\lambda_{Ij}^{XY} = \lambda_{iJ}^{XY} = 0$,还可以考虑对所有的 i 和 j,令 $\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$。不同的统计软件可能会使用不同的约束条件。无论使用哪种约束条件,决定发生比之比的最终公式是唯一的,比如式 8.3 中的 $\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$。

举例来说,假定一个 2×2 表格中的对数发生比之比等于 2.0。按照第一组约束条件,2.0 是 X 的第一个类别和 Y 的第一个类别对应的虚拟变量的乘积的系数。在此情况下,$\lambda_{11}^{XY} = 2.0$,并且 $\lambda_{12}^{XY} = \lambda_{21}^{XY} = \lambda_{22}^{XY} = 0$。在参数总和为零的约束条件下,$\lambda_{11}^{XY} = \lambda_{22}^{XY} = 0.5$, $\lambda_{12}^{XY} = \lambda_{21}^{XY} = -0.5$。无论使用哪一组约束条件,对数发生比之比(式 8.3)都等于 2.0。对于一组参数,与将参数之和设定为 0 相比,将某个基线类别的参数设定为 0 具有一个优点,它允许某些参数的估计值等于无穷大。

8.1.5　关于单元格概率的多项分布模型

以单元格计数的总和 n 为条件,关于$\{\mu_{ij}\}$的泊松分布对数线性模型就成了关于单元格概率 $\{\pi_{ij} = \mu_{ij}/(\sum\sum\mu_{ab})\}$ 的多项分布模型。具体来说,对于饱和模型,

$$\pi_{ij} = \frac{\exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY})}{\sum_a \sum_b \exp(\lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}^{XY})}。 \quad (8.4)$$

这个表达式暗含着有关概率的常见约束条件,即$\{\pi_{ij} \geq 0\}$并且$\sum_i \sum_j \pi_{ij} = 1$。在式8.4多项分布模型中,截距的参数$\lambda$被相互抵消掉。这个参数只与总的样本规模有关,它在泊松分布模型中是随机的,但在多项分布模型中却是固定的。

8.2 关于三维表格的独立性和包括交互项的对数线性模型

在第2.3节中,我们介绍了三维列联表以及相应的数据结构,如条件独立性和同质性关联。可以利用对数线性模型描述三维表格中变量之间的独立性和关联模式。

8.2.1 独立性的类型

关于结果变量X,Y和Z,其三维交叉划分所形成的$I \times J \times K$表格存在着几种可能的独立性情况。我们假定单元格概率$\{\pi_{ijk}\}$服从多项分布,并且$\sum_i \sum_j \sum_k \pi_{ijk} = 1.0$。模型同样适用于均值为$\{\mu_{ijk}\}$的泊松分布样本。

如果对所有的i,j和k,存在

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}, \quad (8.5)$$

我们称这三个变量*相互独立*(*mutually independent*)。对于期望频数$\{\mu_{ijk}\}$,相互独立性所对应的对数线性模型为

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z。 \quad (8.6)$$

如果对所有的i,j和k,存在

$$\pi_{ijk} = \pi_{i+k}\pi_{+j+}, \quad (8.7)$$

我们称变量Y*联合独立于*(*jointly independent*)X和Z。这相当于Y与由X和Z的IK个取值组合所形成的变量之间存在着普通的二维独立性。相应的对数线性模型为

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}。 \quad (8.8)$$

相似地,X也可以联合独立于Y和Z,或者Z联合独立于X和Y。相互独立性(式8.5)意味着,任何一个变量都联合独立于其他变量。

由第2.3节可知,如果在按照Z的取值划分的每个分表中,X和Y都相互独立,我们称给定Z,X和Y之间*条件独立*(*conditional independent*)。换句话说,如果$\pi_{ij|k} = P(X=i, Y=j|Z=k)$,那么对于所有的$i$,$j$和$k$,

$$\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}。$$

等价地,就整个表格的联合概率而言,对于所有的i,j和k,

$$\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}。 \quad (8.9)$$

给定Z,X和Y之间条件独立性的对数线性模型为

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}。 \quad (8.10)$$

条件独立是一种比相互独立和联合独立更弱的情况。相互独立隐含着Y联合独立于X和Z,而后者则又隐含着X和Y之间条件独立。表8.1对这三种类型的独立性进行了总结。

在第2.3.2节中,我们显示了局部关联可能会与边际关联具有很大不同。例如,条件独立性并不意味着边际独立性。当以上所讨论的某种更强的独立性存在时,条件独立性和边际独立性都成立。图8.1解释了这四种独立性之间的关系。

表 8.1 关于对数线性模型独立性的总结

模 型	π_{ijk}的概率形式	对数线性模型中的关联项	解 释
式 8.6	$\pi_{i++}\pi_{+j+}\pi_{++k}$	无	变量间相互独立
式 8.8	$\pi_{i+k}\pi_{+j+}$	λ_{ik}^{XZ}	Y 独立于 X 和 Z
式 8.10	$\pi_{i+k}\pi_{+jk}/\pi_{++k}$	$\lambda_{ik}^{XZ}+\lambda_{jk}^{YZ}$	给定 Z,X 和 Y 独立

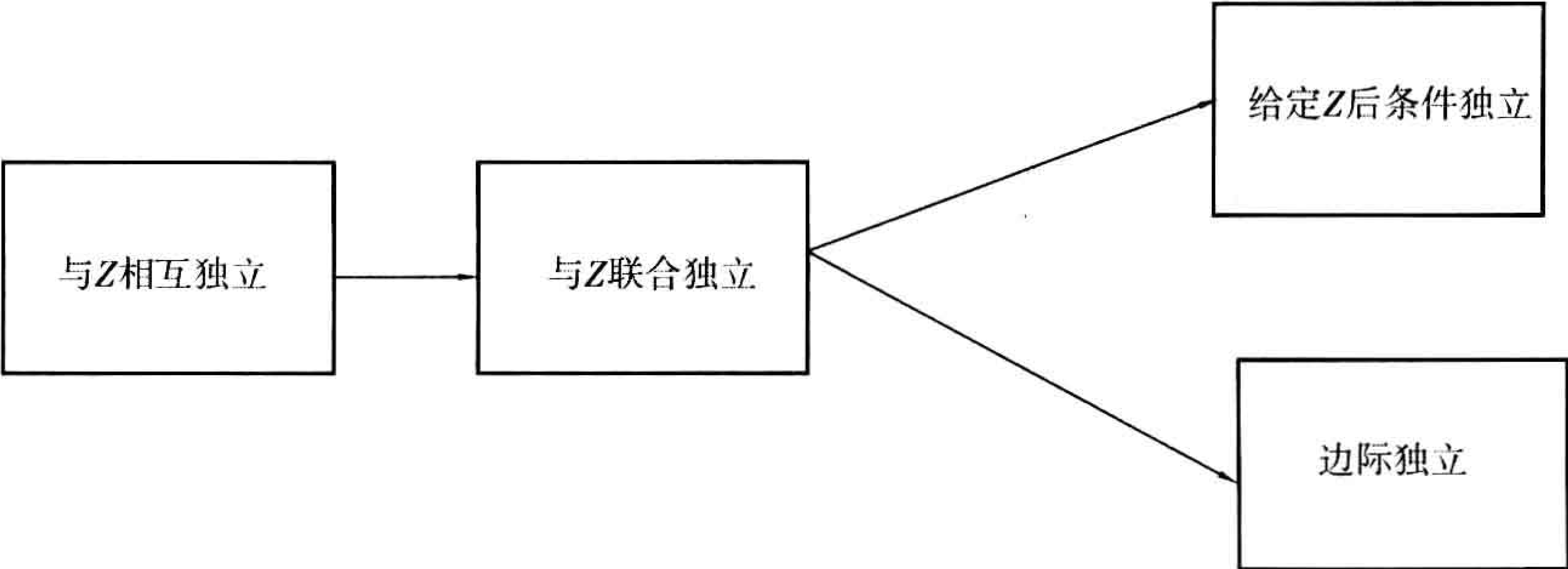


图 8.1 XY 的各种独立性之间的关系

8.2.2 同质性关联与三维交互项

式 8.6、式 8.8 和式 8.10 对数线性模型分别具有三对、两对和一对条件独立的变量。在后两个模型中,包含两个下标的项(比如 λ_{ij}^{XY})对应于条件相依的变量。一个允许所有三对变量都条件相依的模型为

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}。 \quad (8.11)$$

对模型两边求幂,单元格概率具有以下形式:

$$\pi_{ijk} = \psi_{ij}\phi_{jk}\omega_{ik}。$$

除了某些特定情况外,不存在将上述三项表达为$\{\pi_{ijk}\}$的边际形式的封闭解(参见注解 9.2)。

在下一节我们将介绍,对于上述模型,给定第三个变量的每个类别,任意两个变量之间的条件发生比之比都相等。也就是说,每对变量都存在*同质性关联*(*homogeneous association*)(第 2.3.5 节)。式 8.11 模型被称为*同质性关联*或*不存在三维交互项*(*no three-factor interaction*)的对数线性模型。

关于三维表格的一般性对数线性模型可表示为

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}。 \quad (8.12)$$

在虚拟变量的情况下,λ_{ijk}^{XYZ}是关于 X 的第 i 个虚拟变量、Y 的第 j 个虚拟变量以及 Z 的第 k 个虚拟变量之间的乘积项的系数。模型中非冗余参数的总数等于

$$1+(I-1)+(J-1)+(K-1)+(I-1)(J-1)+(I-1)(K-1)+$$
$$(J-1)(K-1)+(I-1)(J-1)(K-1)=IJK,$$

也即单元格计数的总数。这个模型的参数个数与观测值数量一样多,因而是饱和模型。它描述了所有可能的正的$\{\mu_{ijk}\}$。在这种情况下,每对变量之间都可以是条件相依的,并且任意一对变量的发生比之比都可能随着第三个变量的类别变化而变动。

将式 8.12 中的某些参数设定为零,便可得出前面介绍的那些模型。表 8.2 列出了其中的一部分。为了便于区分这些模型,表 8.2 对每个模型都设定了一个符号,由该模型中每个变量的最高阶项组成。例如,关于 X 和 Y 之间条件独立性的模型(式 8.10)对应的符号为(XZ,YZ),因为它的最高阶项是 λ_{ik}^{XZ}和 λ_{jk}^{YZ}。按照在第 6.1 节和第 7.1.2 节中我们对 Logit 模型所使用的标记,上述符号表示的是($X*Z+Y*Z$),它本身则是对既包括

主效应,又包括交互效应的标记$(X+Y+Z+X\times Z+Y\times Z)$的简化。

表 8.2 关于三维表格的对数线性模型

对数线性模型	符号
$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z$	(X,Y,Z)
$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}$	(XY,Z)
$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}+\lambda_{jk}^{YZ}$	(XY,YZ)
$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}+\lambda_{jk}^{YZ}+\lambda_{ik}^{XZ}$	(XY,YZ,XZ)
$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}+\lambda_{jk}^{YZ}+\lambda_{ik}^{XZ}+\lambda_{ijk}^{XYZ}$	(XYZ)

8.2.3 对模型参数的解释

关于对数线性模型参数的解释,关注的是模型的最高阶项。例如,对式 8.11 模型的解释主要是根据二维项来描述条件发生比之比。给定 Z 的取值为 k,X 和 Y 之间的条件关联(*conditional association*)由$(I-1)(J-1)$个发生比之比来描述,如局部发生比之比

$$\theta_{ij(k)}=\frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i,j+1,k}\pi_{i+1,j,k}},\quad 1\leqslant i\leqslant I-1,\quad 1\leqslant j\leqslant J-1 \tag{8.13}$$

类似地,$(I-1)(K-1)$个发生比之比$\{\theta_{i(j)k}\}$描述了 XZ 之间的条件关联,而$(J-1)(K-1)$个发生比之比$\{\theta_{(i)jk}\}$描述了 YZ 之间的条件关联。对数线性模型具有对条件发生比之比设定约束条件的特点。例如,X 和 Y 之间的条件独立性等价于$\{\theta_{ij(k)}=1,i=1,\cdots,I-1,j=1,\cdots,J-1,k=1,\cdots,K\}$。

二维项的参数与条件发生比之比存在直接的关系。具体地说,将模型(XY,XZ,YZ)的式 8.11 代入 $\log\theta_{ij(k)}$得出:

$$\log\theta_{ij(k)}=\log\frac{\mu_{ijk}\mu_{i+1,j+1,k}}{\mu_{i+1,jk}\mu_{i,j+1,k}}=\lambda_{ij}^{XY}+\lambda_{i+1,j+1}^{XY}-\lambda_{i,j+1}^{XY}-\lambda_{i+1,j}^{XY}。 \tag{8.14}$$

对于所有 k,上式右半边都相同,因此,缺乏三维交互项的情况等价于:

$$对于所有的\ i\ 和\ j,\theta_{ij(1)}=\theta_{ij(2)}=\cdots=\theta_{ij(K)}。$$

同样对其他条件发生比之比进行推导,结果表明,模型(XY,XZ,YZ)也等价于:

$$对于所有的\ i\ 和\ k,\theta_{i(1)k}=\theta_{i(2)k}=\cdots=\theta_{i(J)k},$$

以及

$$对于所有的\ j\ 和\ k,\theta_{(1)jk}=\theta_{(2)jk}=\cdots=\theta_{(I)jk}。$$

在任何一个不包括三维交互项的模型中,每对变量都存在同质性关联。

当 X 和 Y 均仅包括两个类别时,只存在一个关于 λ_{ij}^{XY} 的非冗余参数。因此,表达式 8.14可以根据具体的情况加以简化。按照在第 8.1.3 节中关于 2×2 表格的类似推理,在使用虚拟变量并将 X 和 Y 的第二个类别的参数设定为 0 的约束条件下,条件发生比之比简化为 λ_{11}^{XY}。

一般性模型(式 8.12)中的 λ_{ijk}^{XYZ} 项指的是三维交互项。它描述两个变量之间的发生比之比如何随第三个变量取值的变动而变动。我们以 $2\times2\times2$ 表格为例来加以说明。直接代入上述一般性模型的公式,有

$$\begin{aligned}\log\frac{\theta_{11(1)}}{\theta_{11(2)}}&=\log\frac{(\mu_{111}\mu_{221})/(\mu_{121}\mu_{211})}{(\mu_{112}\mu_{222})/(\mu_{122}\mu_{212})}\\&=(\lambda_{111}^{XYZ}+\lambda_{221}^{XYZ}-\lambda_{121}^{XYZ}-\lambda_{211}^{XYZ})-\\&\quad(\lambda_{112}^{XYZ}+\lambda_{222}^{XYZ}-\lambda_{122}^{XYZ}-\lambda_{212}^{XYZ})。\end{aligned}$$

其中仅有一个参数是非冗余的。在将变量第二个类别的参数设定为 0 的约束条件下,这

个发生比之比的对数比等于 λ_{111}^{XYZ}。当 $\lambda_{111}^{XYZ}=0$ 时，$\theta_{11(1)}=\theta_{11(2)}$，即 XY 之间存在同质性关联。

8.2.4 例子：饮酒、抽烟与吸食大麻

表 8.3 所示为一项由莱特州立大学医学院和俄亥俄州代顿联合医学服务中心（Wright State University School of Medicine and the United Health Services in Dayton, Ohio）在 1992 年所进行的一次调查。调查对象是俄亥俄州代顿附近的某一非城区高中毕业班的 2 276 名学生。他们被问到是否曾经有过饮酒、抽烟或者吸食大麻的经历。在这个 $2\times2\times2$ 表格中，由 A 表示饮酒，C 表示抽烟，M 表示吸食大麻。

第 8.7 节将讨论关于对数线性模型的拟合，在这里，我们重点强调如何对模型进行解释。表 8.4 给出了几个对数线性模型的拟合结果。模型（AC,AM,CM）的拟合结果与观察到的数据非常接近，后者也就是饱和模型（ACM）的拟合值。其他模型都拟合得较差。

表 8.3 某高中毕业班学生饮酒、抽烟和吸食大麻的经历

饮酒	抽烟	吸食大麻	
		是	否
是	是	911	538
	否	44	456
否	是	3	43
	否	2	279

来源：由莱特州立大学的 Harry Khamis 友情提供。

表 8.4 关于表 8.3 中数据的对数线性模型拟合值[a]

饮酒	抽烟	吸食大麻	对数线性模型				
			(A,C,M)	(AC,M)	(AM,CM)	(AC,AM,CM)	(ACM)
是	是	是	540.0	611.2	909.24	910.4	911
		否	740.2	837.8	438.84	538.6	538
	否	是	282.1	210.9	45.76	44.6	44
		否	368.7	289.1	555.16	455.4	456
否	是	是	90.6	19.4	4.76	3.6	3
		否	124.2	26.6	142.16	42.4	43
	否	是	47.3	118.5	0.24	1.4	2
		否	64.9	162.5	179.84	279.6	279

a A，饮酒；C，抽烟；M，吸食大麻。

表 8.5 通过计算条件发生比之比和边际发生比之比的估计值，展示了模型的关联模式。例如，在 AC 条件独立性模型（AM,CM）中，AC 条件关联的值 1.0 就是在 M 的两个取值水平上所拟合的 AC 发生比之比的共同值，

$$1.0=\frac{909.24\times0.24}{45.76\times4.76}=\frac{438.84\times179.84}{555.16\times142.16}。$$

同一个模型中，AC 边际关联的值 2.7 是所拟合的 AC 边际表的发生比之比。观察数据的发生比之比是由饱和模型（ACM）的相应结果计算而得。

表 8.5 表明，对于每个未包括在模型中的成对项，所估计的发生比之比都等于 1.0，比如模型（AM,CM）中的 AC 项。在此模型中，所估计的 AC 边际发生比之比并不等于

1.0,因为条件独立性并不意味着边际独立性。有些模型会要求条件关联必须与相应的边际关联相等。我们将在第9.1.2节介绍这样一种情况。

表8.5 表8.4中的对数线性模型的发生比之比估计值

模 型	条件关联			边际关联		
	AC	*AM*	*CM*	*AC*	*AM*	*CM*
(*A*,*C*,*M*)	1.0	1.0	1.0	1.0	1.0	1.0
(*AC*,*M*)	17.7	1.0	1.0	17.7	1.0	1.0
(*AM*,*CM*)	1.0	61.9	25.1	2.7	61.9	25.1
(*AC*,*AM*,*CM*)	7.8	19.8	17.3	17.7	61.9	25.1
(*ACM*)类别1	13.8	24.3	17.5	17.7	61.9	25.1
(*ACM*)类别2	7.7	13.5	9.7			

模型(*AC*,*AM*,*CM*)允许存在所有的成对关联,但却要求两个变量在第三个变量的任意取值上具有同质性发生比之比。这个模型所拟合的关于*AC*的条件发生比之比等于7.8。这个发生比之比可以利用在*M*的每个取值上的模型拟合值或由式8.14利用$\exp(\hat{\lambda}_{11}^{AC}+\hat{\lambda}_{22}^{AC}-\hat{\lambda}_{12}^{AC}-\hat{\lambda}_{21}^{AC})$来计算。

表8.5的结果表明,关于发生比之比的估计在很大程度上依赖于模型本身。这种情况突出反映了选择一个好的模型的重要性。只有当模型在一定程度上拟合得很好时,相应的估计值才有意义。在下一节中,我们讨论拟合优度的问题。

8.3 对数线性模型的统计推断

一个拟合结果很好的对数线性模型,是描述分类结果变量之间的关联并进行统计推断的基础。在这里,关于模型参数的拟合检验和统计推断的标准方法仍然适用。

8.3.1 卡方拟合优度检验

跟通常一样,X^2和G^2通过比较单元格的拟合值与实际观察到的计数来检验模型是否成立。这里,自由度df等于单元格计数的数量减去模型的参数个数。

表8.6 关于表8.4中的对数线性模型的拟合优度检验

模 型	G^2	X^2	df	*P*值[a]
(*A*,*C*,*M*)	1 286.0	1 411.4	4	<0.001
(*A*,*CM*)	534.2	505.6	3	<0.001
(*C*,*AM*)	939.6	824.2	3	<0.001
(*M*,*AC*)	843.8	704.9	3	<0.001
(*AC*,*AM*)	497.4	443.8	2	<0.001
(*AC*,*CM*)	92.0	80.8	2	<0.001
(*AM*,*CM*)	187.8	177.6	2	<0.001
(*AC*,*AM*,*CM*)	0.4	0.4	1	0.54
(*ACM*)	0.0	0.0	0	—

a G^2统计量的*P*值。

对于有关高中生的调查数据(表8.3),表8.6给出了对几个对数线性模型拟合结果的检验情况。在这些模型中,不包括任何关联项的模型拟合得很差,包括所有三个成对

关联项的模型(AC,AM,CM)则拟合得很好($P=0.54$)。其他检验标准也给出了相同的结论,比如最小化

$$\mathrm{AIC}=-2(\text{最大对数似然值}-\text{模型中的参数个数}),$$

或者等价地,最小化$[G^2-2(\mathrm{df})]$。

8.3.2 条件关联的统计推断

通过比较不同的对数线性模型,可以用来检验是否存在条件关联。在这里,似然比统计量 $-2(L_0-L_1)$ 与 $G^2(M_0|M_1)=G^2(M_0)-G^2(M_1)$ 完全一致,即不包括某项与包括该项的模型之间的偏离度之差。在模型(XY,XZ,YZ)中,考虑 XY 之间条件独立性的假设,这时,对于 $(I-1)(J-1)$ 个 XY 的关联参数有 $H_0:\lambda_{ij}^{XY}=0$。检验统计量为 $G^2(XZ,YZ)-G^2(XY,XZ,YZ)$,自由度 $\mathrm{df}=(I-1)(J-1)$。这与在第 7.5 节所介绍的关于定类变量的广义 CMH 检验以及相应基于模型的检验具有相同的目的。

举例来说,关于饮酒和抽烟的条件独立性检验将模型(AM,CM)与备择模型(AC,AM,CM)相比较。检验统计量等于

$$G^2[(AM,CM)\mid(AC,AM,CM)]=187.8-0.4=187.4,$$

它的自由度 $\mathrm{df}=2-1=1$($P<0.001$)。类似地,将(AC,CM)和(AC,AM)与(AC,AM,CM)进行比较的统计量也给出了强有力的证据表明,存在关于 AM 和 CM 的条件关联,所以我们利用模型(AC,AM,CM)对表 8.3 展开进一步分析。

表 8.7 对表 8.3 数据拟合对数线性模型的输出结果

Criteria For Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	1	0.374 0	0.374 0
Pearson Chi-Square	1	0.401 1	0.401 1

Parameter			Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept			5.633 4	0.059 7	8 903.96	<.000 1
a	1		0.487 7	0.075 8	41.44	<.000 1
c	1		-1.886 7	0.162 7	134.47	<.000 1
m	1		-5.309 0	0.475 2	124.82	<.000 1
a×m	1	1	2.986 0	0.464 7	41.29	<.000 1
a×c	1	1	2.054 5	0.174 1	139.32	<.000 1
c×m	1	1	2.847 9	0.163 8	302.14	<.000 1

LR Statistics

Source	DF	Chi-Square	Pr > ChiSq
a×m	1	91.64	<.000 1
a×c	1	187.38	<.000 1
c×m	1	497.00	<.000 1

在大样本的情况下,一些统计上显著的效应可能会很弱,而且并不具有实质意义,更值得关注的问题是,关联本身是否足够强大而具有实际意义。这时候,置信区间比统计检验更有价值。表 8.7 展示了模型(AC,AM,CM)的拟合结果,利用(1,0)虚拟变量代表每个变量的类别划分,并将最后一行和最后一列的参数设定为零。在假定模型(AC,AM,CM)成立的情况下,考虑关于 AC 的条件发生比之比。表 8.7 报告的结果为 $\hat{\lambda}_{11}^{AC}=2.054$,$\mathrm{SE}=0.174$。在上述的约束条件下,这就是所估计的条件对数发生比之比。关于 AC 的真

实条件发生比之比的 95% 的沃尔德置信区间为 exp[2.054 ± 1.96(0.174)]，即(5.5，11.0)。无论是否吸食过大麻，在抽烟与饮酒之间存在着很强的正向关联。

在模型(*AC*,*AM*,*CM*)中，关于 *AM* 的条件发生比之比的 95% 的沃尔德置信区间为(8.0,49.2)，关于 *CM* 的相应区间为(12.5,23.8)。这些区间都很宽，但是关联本身都很强。表 8.5 显示，边际关联的估计值甚至更强，也即，控制其中一个结果变量会在某种程度上缓和另外两个变量之间的关联。

本节中的分析主要强调变量之间的关联。一种不同的分析思路是强调比较单个变量的边际分布，如考虑抽烟的学生是否比饮酒或吸食大麻的学生多。这样的分析，我们将留在第 10.1 节中介绍。

8.4 更高维数的对数线性模型

三维表格的对数线性模型中，由于可能存在的各种关联项更多，因而比二维表格的情况更为复杂。不过，关于三维表格的对数线性模型可以很容易地扩展到多维表格的情况。随着维数的增加，也会出现一些复杂的问题。一方面，可能的关联和交互项数量上升，这使得模型选择变得更困难。另一方面，单元格的数量也会大幅增加，这可能会导致估计值存在与否以及渐近理论是否适用的问题，正如我们将在第 9.8 节所讨论的那样。

8.4.1 四维列联表

我们通过一个包括变量 *W*,*X*,*Y* 和 *Z* 的四维表格来展示更高维度的对数线性模型。当模型不包括三维交互项时，对它的解释相对简单。相应的模型是以下模型的特例：

$$\log \mu_{hijk} = \lambda + \lambda_h^W + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hi}^{WX} + \lambda_{hj}^{WY} + \lambda_{hk}^{WZ} + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

该模型可表示为(*WX*,*WY*,*WZ*,*XY*,*XZ*,*YZ*)。每对变量之间都存在条件相依，但是在另外两个变量的每个取值组合上，它们的发生比之比是相同的。不包括某两个变量之间的二维项的模型则意味着，在给定其他两个变量后，这两个变量条件独立。

在四维表格的情况下，很多模型包括三维交互项，它们可以是 *WXY*,*WXZ*,*WYZ* 或 *XYZ* 中的任意项。对于模型(*WXY*,*WZ*,*XZ*,*YZ*)，每对变量都是条件相依的，但是在 *Z* 的每个取值水平上的 *WX* 关联、*WY* 关联和 *XY* 关联可能会随着另外一个变量取值的变动而变动，而 *Z* 和其他三个变量之间的条件关联则是同质的。相应的饱和模型在包括所有上述三维项的同时，再加上一个四维交互项。

8.4.2 例子：车祸的数据

表 8.8 所示为 1991 年缅因州发生的涉及小汽车和轻型卡车的车祸中 68 694 名乘客的状况。该表按照性别(*G*)、车祸地点(*L*)、是否使用安全带(*S*)，以及是否受伤(*I*)对数据进行了划分。表 8.8 还给出了受伤乘客的样本比例。在 *GL* 的每一个取值组合点，系安全带的乘客的受伤比例都减少了大约一半。

表 8.9 给出了几个对数线性模型的拟合检验结果。为了考察模型所需要的复杂程度，我们比较具有不同复杂程度的模型(*G*,*I*,*L*,*S*)、(*GI*,*GL*,*GS*,*IL*,*IS*,*LS*)、以及(*GIL*,*GIS*,*GLS*,*ILS*)。在这些模型中，相互独立性模型(*G*,*I*,*L*,*S*)的拟合结果很差。模型(*GI*,*GL*,*GS*,*IL*,*IS*,*LS*)的拟合结果比(*G*,*I*,*L*,*S*)好得多，但仍然不够充分($P<0.001$)。模型(*GIL*,*GIS*,*GLS*,*ILS*)对数据拟合得很好($G^2=1.3$，df = 1)，但是它很复杂而且难以解释。

这些结果表明，我们应考虑那些比（GI,GL,GS,IL,IS,LS）复杂但却比（GIL,GIS,GLS,ILS）简单的模型。

表 8.8 关于受伤情况、使用安全带、性别和车祸地点的对数线性模型[a]

性别	事故地点	是否系安全带	是否受伤		(GI,GL,GS,IL,IS,LS)		(GLS,GI,IL,IS)		受伤的样本比例
			否	是	否	是	否	是	
女性	城市	否	7 287	996	7 166.4	993.0	7 273.2	1 009.8	0.12
		是	11 587	759	11 748.3	721.3	11 632.6	713.4	0.06
	农村	否	3 246	973	3 353.8	988.8	3 254.7	964.3	0.23
		是	6 314	757	5 985.5	781.9	6 093.5	797.5	0.11
男性	城市	否	10 381	812	10 471.5	845.1	10 358.9	834.1	0.07
		是	10 969	380	10 837.8	387.6	10 959.2	389.8	0.03
	农村	否	6 123	1 084	6 045.3	1 038.1	6 150.2	1 056.8	0.15
		是	6 693	513	6 811.4	518.2	6 697.6	508.4	0.07

a G，性别；I，是否受伤；L，车祸地点；S，是否系安全带。

来源：由缅因州奥古斯特医疗发展中心的 *Cristanna Cook* 友情提供。

表 8.9 关于表 8.8 中的对数线性模型的拟合优度检验

模型	G^2	df	p 值
(G,I,L,S)	2 792.8	11	<0.000 1
(GI,GL,GS,IL,IS,LS)	23.4	5	<0.001
(GIL,GIS,GLS,ILS)	1.3	1	0.25
(GIL,GS,IS,LS)	18.6	4	0.001
(GIS,GL,IL,LS)	22.8	4	<0.001
(GLS,GI,IL,IS)	7.5	4	0.11
(ILS,GI,GL,GS)	20.6	4	<0.001

不过，我们首先还是来分析一下模型（GI,GL,GS,IL,IS,LS），它主要关注成对的关联。表 8.8 给出了相应的拟合值。表 8.10 报告了模型所估计的条件发生比之比。读者可以直接利用在其中两个变量的任意取值组合上另外两个变量的分表的拟合值来计算这些结果。也可以通过模型的参数估计值来直接计算，比如，$0.44=\exp(\hat{\lambda}_{11}^{IS}+\hat{\lambda}_{22}^{IS}-\hat{\lambda}_{12}^{IS}-\hat{\lambda}_{21}^{IS})$。

表 8.10 利用表 8.8 中的模型所估计的条件发生比之比

发生比之比	对数线性模型	
	(GI,GL,GS,IL,IS,LS)	(GLS,GI,IL,IS)
GI	0.58	0.58
IL	2.13	2.13
IS	0.44	0.44
GL S = 否	1.23	1.33
S = 是	1.23	1.17
GS L = 城市	0.63	0.66
L = 农村	0.63	0.58
LS G = 女性	1.09	1.17
G = 男性	1.09	1.03

由于数据的样本规模很大，模型关于发生比之比的估计非常精确。例如，所估计的 IS 条件对数发生比之比 -0.814 的标准误为 0.028。相应的真实发生比之比的 95% 的沃尔德置信区间是 $\exp[-0.814\pm1.96(0.028)]$，或（0.42，0.47）。这个模型估计结果显

示,在性别-车祸地点的每个取值组合上,系安全带的乘客受伤的发生比还不到没有系安全带的乘客的一半。表 8.10 中拟合的发生比之比还表明,在给定其他因素后,发生在农村的车祸受伤的可能性比城市更大,女性受伤的可能性大于男性。根据该模型,所估计的系安全带的男性受伤的发生比仅仅是女性的 0.63 倍。

包括三维交互项的模型解释起来更加复杂。表 8.9 显示了在模型(GI,GL,GS,IL,IS,LS)中加入一个三维交互项的结果。在所有的四个可能模型中,(GLS,GI,IL,IS)拟合得最好。表 8.8 也给出了它的拟合结果。由于样本规模很大,模型的 G^2 值表明该模型拟合得非常好。

在模型(GLS,GI,IL,IS)中,每一对变量都是条件相依的,并且在 I 的每个类别中,其他任意两个变量之间的关联都随着另一个变量取值的变动而变动。就这个模型来说,只是解释 GL,GS 和 LS 这些二维项是不恰当的。由于 I 没有出现在任何三维项中,I 和每个变量的条件发生比之比(参见表 8.10 的上半部分)在其他两个变量的每个取值组合上都是相同的。

当一个模型包括三维交互项但不包括更高阶的项时,大家可以通过计算两个变量在第三个变量的每个取值水平上所拟合的发生比之比来分析这个交互项。对于交互项所不包括的变量,它取任意值都不影响计算的结果。表 8.10 的下半部分通过模型(GLS,GI,IL,IS)表明了这一点。例如,对于"L = 城区",GS 发生比的拟合值 0.66 参考的是发生在城区的车祸的四个拟合值,这些值既可以是"没有受伤"的四个值,也可以是"受伤"的四个值;比如,$0.66 = (7\ 273.2 \times 10\ 959.2)/(11\ 632.6 \times 10\ 358.9)$。

8.4.3 大样本、统计显著性与现实显著性

模型(GLS,GI,IL,IS)的拟合结果看上去比模型(GI,GL,GS,IL,IS,LS)好得多。二者的 G^2 值之差为 $23.4 - 7.5 = 15.9$,相应的自由度为 $\mathrm{df} = 5 - 4 = 1$($P = 0.000\ 1$)。然而,表 8.10 显示,三维交互项的强度很弱。G,L 和 S 中的任意两个变量之间所拟合的发生比之比在第三个变量的两个取值水平上很相似。模型(GLS,GI,IL,IS)所对应的拟合结果的显著改善,主要是反映了巨大的样本规模的影响。

对于所有的检验,一个在统计上显著的效应并不一定具有现实意义。当样本规模特别大时,应当将注意力集中在参数估计而不是假设检验上。例如,对比表 8.10 中两个模型所拟合的发生比之比显示,在大多数情况下,较简单的模型(GI,GL,GS,IL,IS,LS)就足够了。

8.4.4 差异指数

对于一个任意维度的表格,其单元格计数为 $\{n_i = np_i\}$,而相应的拟合值为 $\{\hat{\mu}_i = n\hat{\pi}_i\}$,可以通过下面的差异指数(*dissimilarity index*)(Gini,1914)来反映模型对数据拟合的接近程度:

$$\hat{\Delta} = \sum_i |n_i - \hat{\mu}_i|/2n = \sum_i |p_i - \hat{\pi}_i|/2。$$

差异指数的取值范围为 0 到 1,值越小表示拟合得越好。它代表要使模型能够完全地拟合数据,需要将样本中的对象移动到不同的单元格中去的比例。

差异指数 $\hat{\Delta}$ 是对描述模型拟合不足的相应总体指数 Δ 的估计。当模型完全成立时,存在 $\Delta = 0$。在实际应用中,对于非饱和模型来说,总会有 $\Delta > 0$。估计值 $\hat{\Delta}$ 有助于分析所出现的拟合不足是否具有重要的现实意义。当 $\hat{\Delta} < 0.02$ 或 0.03 时,样本数据与模型所

描述的模式相当接近，尽管模型本身并不完美。当 Δ 接近于 0 时，$\hat{\Delta}$ 一般会高估 Δ，尤其是在小样本的情况下。Firth 和 Kuha(2000) 给出了关于 $\hat{\Delta}$ 的近似方差，并讨论了降低其估计偏差的办法。

在表 8.8 中，模型(GI,GL,GS,IL,IS,LS)的差异指数为 $\hat{\Delta}=0.008$，而模型(GLS,GI,IL,IS)的差异指数为 $\hat{\Delta}=0.003$。对于上述任一个模型来说，只需要移动不到 1% 的数据就能得到完全拟合。模型(GI,GL,GS,IL,IS,LS)所对应的较大 G^2 值表明，这个模型并不真正成立。然而，较小的 $\hat{\Delta}$ 值却表明，从现实的角度来说，它的拟合是可以接受的。

8.5 对数线性模型与 Logit 模型的关系

对数线性模型将分类的结果变量视为对称的，主要关注变量的联合分布中的关联和交互效应。相反，Logit 模型描述单个分类结果变量如何取决于其他解释变量。从模型的类型来看，二者似乎是明显不同的，但是它们之间却存在着密切的联系。一方面，在对数线性模型中，通过构建其中一个结果变量的 Logit 有助于解释模型结果。另一方面，解释变量为分类变量的 Logit 模型具有与其等价的对数线性模型。

8.5.1 利用 Logit 模型来解释对数线性模型

为了理解对数线性模型公式的含义，构建关于其中一个变量的 Logit 很有帮助。我们以模型(XY,XZ,YZ)为例来加以说明。当 Y 是二分变量时，它的 Logit 等于

$$\begin{aligned}\log\frac{P(Y=1\mid X=i,Z=k)}{P(Y=2\mid X=i,Z=k)}&=\log\frac{\mu_{i1k}}{\mu_{i2k}}=\log\mu_{i1k}-\log\mu_{i2k}\\&=(\lambda+\lambda_i^X+\lambda_1^Y+\lambda_k^Z+\lambda_{i1}^{XY}+\lambda_{ik}^{XZ}+\lambda_{1k}^{YZ})\\&\quad-(\lambda+\lambda_i^X+\lambda_2^Y+\lambda_k^Z+\lambda_{i2}^{XY}+\lambda_{ik}^{XZ}+\lambda_{2k}^{YZ})\\&=(\lambda_1^Y-\lambda_2^Y)+(\lambda_{i1}^{XY}-\lambda_{i2}^{XY})+(\lambda_{1k}^{YZ}-\lambda_{2k}^{YZ})。\end{aligned}$$

第一个括号中的项是一个常数，它不取决于 i 或 k。第二个括号中的项取决于 X 的类别 i。第三个括号中的项取决于 Z 的类别 k。这个 Logit 具有以下的可加形式：

$$\text{Logit}[P(Y=1)\mid X=i,Z=k]=\alpha+\beta_i^X+\beta_k^Z。\tag{8.15}$$

按照利用模型的预测变量来标示 Logit 模型的做法，我们将这个模型表示为($X+Z$)。

在第 5.4.1 节中，我们介绍了这个 Logit 模型。当 Y 是二分变量时，对数线性模型(XY,XZ,YZ)与之等价。对数线性模型中解释变量之间的关联项 λ_{ik}^{XZ} 在 Logit 所定义的对数之差中相互抵消。这个 Logit 模型并不分析该关联项。

8.5.2 例子：再析车祸的数据

在第 8.4.2 节我们指出，对于缅因州车祸的数据(表 8.8)，对数线性模型(GLS,GI,LI,IS)

$$\begin{aligned}\log\mu_{gils}=\lambda&+\lambda_g^G+\lambda_i^I+\lambda_l^L+\lambda_s^S+\lambda_{gi}^{GI}+\lambda_{gl}^{GL}+\lambda_{gs}^{GS}+\\&\lambda_{il}^{IL}+\lambda_{is}^{IS}+\lambda_{ls}^{LS}+\lambda_{gls}^{GLS}\end{aligned}$$

拟合得很好。这里，我们自然地将是否受伤(I)作为结果变量，并将性别(G)、地点(L)，以及是否使用安全带(S)作为解释变量；或者，也可以将 S 作为结果变量，而用 G 和 L 作为解释变量。读者可以证明，这个对数线性模型等价于 Logit 模型($G+L+S$)，

$$\text{Logit}[P(I=1)\mid G=g, L=l, S=s]=\alpha+\beta_g^C+\beta_l^L+\beta_s^S。\quad (8.16)$$

例如,两个模型中是否系安全带的效应满足 $\beta_s^S=\lambda_{1s}^{IS}-\lambda_{2s}^{IS}$。在计算 Logit 的过程中,对数线性模型里所有不包含关于受伤的标示 i 的项都相互抵消。上述 Logit 模型的拟合值、拟合优度统计量、残差自由度,以及标准化皮尔逊残差与对数线性模型都完全相同。

用来描述对 I 的效应的发生比之比对应于对数线性模型中二维项的参数以及 Logit 模型中主效应的参数。在 Logit 模型中,S 对 I 的效应的对数发生比之比等于 $\beta_1^S-\beta_2^S$,这相当于对数线性模型中的 $\lambda_{11}^{IS}+\lambda_{22}^{IS}-\lambda_{12}^{IS}-\lambda_{21}^{IS}$。不论统计软件怎么设定约束条件,两个模型的估计结果都相同。对于表 8.8 的数据,Logit 模型的结果为 $\hat{\beta}_1^S-\hat{\beta}_2^S=-0.817$,而对数线性模型的结果也是 $\hat{\lambda}_{11}^{IS}+\hat{\lambda}_{22}^{IS}-\hat{\lambda}_{12}^{IS}-\hat{\lambda}_{21}^{IS}=-0.817$。

对数线性模型是将表 8.8 中的 16 个单元格计数视为 16 个独立的泊松分布变量的广义线性模型。Logit 模型则是将整个表格作为二项分布计数的广义线性模型。以 I 作为结果变量的 Logit 模型把 GLS 表的边际 $\{n_{g+ls}\}$ 视为给定的,并把 $\{n_{g1ls}\}$ 作为关于该结果变量的 8 个独立的二项分布变量。尽管对数线性模型与 Logit 模型背后的抽样模型不同,它们的拟合结果完全一致。

8.5.3 对数线性模型和 Logit 模型之间的对应关系

在从对数线性模型(XY,XZ,YZ)推导 Logit 模型($X+Z$)的过程中(参见式 8.15),λ_{ik}^{XZ} 项被抵消掉了。因而,不包括这项的对数线性模型(XY,YZ)似乎应该也等价于上述 Logit 模型。确实,根据模型(XY,YZ)构建关于 Y 的 Logit 也会推出相同的 Logit 公式。但是,与 Logit 模型具有相同拟合结果的对数线性模型包括一个描述解释变量之间关系的交互项。Logit 模型对解释变量之间的关系不做任何假定,因而它允许在解释变量之间存在任意形式的交互效应。

表 8.11 列出了在三维表格中当 Y 是二分结果变量时,等价的 Logit 模型和对数线性模型。每个对数线性模型都包括了在 Logit 模型中描述解释变量之间关系的 XZ 关联项。简单的对数线性模型(Y,XZ)表明,Y 联合独立于 X 和 Z,它等价于只具有截距项的 Logit 模型。饱和的对数线性模型(XYZ)包含了三维交互项,当 Y 是二分结果变量时,这个模型等价于一个包括预测变量 X 和 Z 之间交互项的 Logit 模型。例如,X 对 Y 的效应取决于 Z,意味着 XY 的发生比之比随 Z 的取值的变动而变动。这个 Logit 模型也是饱和模型。

表 8.11 在三维表格中关于二分结果变量 Y 的等价的对数线性模型和 Logit 模型

对数线性模型	Logit 模型	Logit 标示符号
(Y,XZ)	α	(—)
(XY,XZ)	$\alpha+\beta_i^X$	(X)
(YZ,XZ)	$\alpha+\beta_k^Z$	(Z)
(XY,YZ,XZ)	$\alpha+\beta_i^X+\beta_k^Z$	($X+Z$)
(XYZ)	$\alpha+\beta_i^X+\beta_k^Z+\beta_{ik}^{XZ}$	($X\times Z$)

当 Y 具有多个类别时,基线类别 Logit 模型与对数线性模型之间也存在类似的对应关系。对数线性模型的一个优点在于它的普适性。当存在多个结果变量时,仍然可以使用对数线性模型。例如,在第 8.2.4 节关于饮酒-抽烟-吸食大麻的例子中,我们利用对数线性模型研究三个结果变量之间的关联模式。当至少有两个变量为结果变量时,对数线

性模型是最自然的选择。当只有一个变量是结果变量时，直接使用 Logit 模型更合理一些。

8.5.4 广义对数线性模型*

令 $\mathbf{n}=(n_1,\cdots,n_N)'$ 和 $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_N)'$ 分别表示由一个列联表中 N 个单元格的观察计数和期望计数所组成的列向量，其中 $n=\sum_i n_i$。为简便起见，我们在这里只使用一个下标，但表格本身可以是多维的。对于均值为正的泊松分布，对数线性模型具有以下形式：

$$\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{8.17}$$

其中 $\mathbf{X}$ 为模型矩阵，列向量 $\boldsymbol{\beta}$ 为模型参数。

我们通过关于 2×2 表格的独立性模型 $\log \mu_{ij}=\lambda+\lambda_i^X+\lambda_j^Y$ 来加以说明。在 $\lambda_2^X=\lambda_2^Y=0$ 的约束条件下，该模型为

$$\begin{bmatrix}\log \mu_{11}\\ \log \mu_{12}\\ \log \mu_{21}\\ \log \mu_{22}\end{bmatrix}=\begin{bmatrix}1&1&1\\1&1&0\\1&0&1\\1&0&0\end{bmatrix}\begin{bmatrix}\lambda\\ \lambda_1^X\\ \lambda_1^Y\end{bmatrix}。$$

有关式 8.17 的扩展可以囊括许多其他的模型。此时，扩展形式的广义*对数线性模型*（*generalized longlinear model*）可表示为，对于矩阵 $\mathbf{C}$ 和 $\mathbf{A}$，

$$\mathbf{C}\log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}。 \tag{8.18}$$

当 $\mathbf{C}$ 和 $\mathbf{A}$ 是单位矩阵时，它简化为普通的对数线性模型式 8.17。关于二分结果变量和多类别结果变量的 Logit 模型也都是它的特例。

例如，关于 2×2 表格的独立性对数线性模型等价于对于 X 的每一行，Y 的 Logit 都相等的模型（参见第 8.1.2 节）。这个 Logit 模型具有式 8.18 的形式：$\mathbf{A}$ 是一个 4×4 的单位矩阵，因而 $\mathbf{A}\boldsymbol{\mu}$ 是一个 4×1 的向量 $\boldsymbol{\mu}=(\mu_{11},\mu_{12},\mu_{21},\mu_{22})'$；利用

$$\mathbf{C}=\begin{bmatrix}1&-1&0&0\\0&0&1&-1\end{bmatrix},$$

$\mathbf{C}\log(\mathbf{A}\boldsymbol{\mu})$ 的乘积形成了第 1 行和第 2 行的 Logit；这时，$\mathbf{X}=(1,1)'$ 是一个 2×1 矩阵，$\boldsymbol{\beta}$ 等于常数 α，因而 $\mathbf{X}\boldsymbol{\beta}$ 构成了这两个 Logit 的一个共同值。

在第 10 章和第 11 章中，我们将利用广义对数线性模型来讨论那些不属于目前所定义的广义线性模型范畴的其他模型。其中的一个例子是对多元结果变量的边际分布进行的模型分析。

8.6 对数线性模型的拟合：似然方程和渐近分布*

在讨论对数线性模型的拟合之前，我们首先推导充分统计量和似然方程。接下来，我们介绍有关模型参数和单元格概率的最大似然估计值的大样本正态分布。最后，我们通过对三维表格的模型分析来加以展示。为简便起见，这里的推导过程以泊松分布样本的模型为例，这种模型不像多项分布样本那样要求对参数加以限定。

8.6.1 最小充分统计量

在三维表格中，单元格计数 $\{Y_{ijk}=n_{ijk}\}$ 的联合泊松分布概率等于

$$\prod_i \prod_j \prod_k \frac{e^{-\mu_{ijk}} \mu_{ijk}^{n_{ijk}}}{n_{ijk}!},$$

其中的乘积项是对表格中所有单元格的乘积。对数似然函数的核函数为

$$L(\boldsymbol{\mu}) = \sum_i \sum_j \sum_k n_{ijk} \log \mu_{ijk} - \sum_i \sum_j \sum_k \mu_{ijk}。 \tag{8.19}$$

对于一般的对数线性模型式 8.12,该式简化为

$$\begin{aligned} L(\boldsymbol{\mu}) = n\lambda + \sum_i n_{i++}\lambda_i^X + \sum_j {}_{n+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z + \\ \sum_i \sum_j n_{ij+}\lambda_{ij}^{XY} + \sum_i \sum_k n_{i+k}\lambda_{ik}^{XZ} + \sum_j \sum_k n_{+jk}\lambda_{jk}^{YZ} + \\ \sum_i \sum_j \sum_k n_{ijk}\lambda_{ijk}^{XYZ} - \sum_i \sum_j \sum_k \exp(\lambda + \cdots + \lambda_{ijk}^{XYZ}) \end{aligned} \tag{8.20}$$

由于泊松分布属于指数分布族,参数的系数就是它的充分统计量。对于这个饱和模型,$\{n_{ijk}\}$ 是 $\{\lambda_{ijk}^{XYZ}\}$ 的系数,因而这里不存在对数据的简化。对于较简单的模型,某些参数为 0,相应地可以简化式 8.20。例如,在相互独立性的模型 $\{X,Y,Z\}$ 中,充分统计量就是式 8.20 中 $\{\lambda_i^X\}$,$\{\lambda_j^Y\}$,以及 $\{\lambda_k^Z\}$ 的系数。它们分别为 $\{n_{i++}\}$、$\{n_{+j+}\}$ 和 $\{n_{++k}\}$。

表 8.12 列出了几个对数线性模型的最小充分统计量。它们都是变量所属的最高阶项的系数。事实上,这些就是出现在模型标示符号中的各项的边际分布。较简单的模型是对样本信息更概括的使用。例如,模型 $\{X,Y,Z\}$ 仅使用了单维的边际分布,而模型 (XY,XZ,YZ) 则使用了相应的二维边际表格。

表 8.12 拟合对数线性模型的最小充分统计量

模　型	最小充分统计量
(X,Y,Z)	$\{n_{i++}\}$, $\{n_{+j+}\}$, $\{n_{++k}\}$
(XY,Z)	$\{n_{ij+}\}$, $\{n_{++k}\}$
(XY,YZ)	$\{n_{ij+}\}$, $\{n_{+jk}\}$
(XY,XZ,YZ)	$\{n_{ij+}\}$, $\{n_{i+k}\}$, $\{n_{+jk}\}$

8.6.2 对数线性模型的似然方程

模型的拟合值就是相应似然方程的解。我们利用对数线性模型的一般性表达式 8.17来推导似然方程。对于具有 $\boldsymbol{\mu} = E(\mathbf{n})$ 的计数向量 $\mathbf{n}$,模型为 $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$,其中对于所有的 i,都存在 $\log(\mu_i) = \sum_j x_{ij}\beta_j$。

对式 8.19 进行扩展,关于泊松分布样本的对数似然函数为

$$\begin{aligned} L(\boldsymbol{\mu}) &= \sum_i n_i \log \mu_i - \sum_i \mu_i \\ &= \sum_i n_i \left(\sum_j x_{ij}\beta_j \right) - \sum_i \exp\left(\sum_j x_{ij}\beta_j \right)。 \end{aligned} \tag{8.21}$$

其中 β_j 的充分统计量是它的系数,即 $\sum_i n_i x_{ij}$。由于

$$\frac{\partial}{\partial \beta_j}\left[\exp\left(\sum_j x_{ij}\beta_j \right) \right] = x_{ij}\exp\left(\sum_j x_{ij}\beta_j \right) = x_{ij}\mu_i,$$

$$\frac{\partial L(\boldsymbol{\mu})}{\partial \beta_j} = \sum_i n_i x_{ij} - \sum_i \mu_i x_{ij}, j = 1,2,\cdots,p。$$

令这些导数等于零,便得出似然方程组。似然方程的形式为

$$\mathbf{X}'\mathbf{n} = \mathbf{X}'\hat{\boldsymbol{\mu}}。 \tag{8.22}$$

这些方程式令充分统计量等于它们的期望值,这一结果在式 4.29 中已根据广义线性模型理论推导得出。对于到目前为止本章所讨论的模型而言,充分统计量就是模型标示符号中的边际表。

举例来说,考虑模型(XZ,YZ),它的对数似然函数为令 $\lambda^{XY}=\lambda^{XYZ}=0$ 的式 8.20。对其求导可得:

$$\frac{\partial L}{\partial \lambda_{ik}^{XZ}} = n_{i+k} - \mu_{i+k} \quad 和 \quad \frac{\partial L}{\partial \lambda_{jk}^{YZ}} = n_{+jk} - \mu_{+jk},$$

由此得出似然方程组:

$$对所有的 i 和 k, \ \hat{\mu}_{i+k} = n_{i+k}, \tag{8.23}$$

$$对所有的 j 和 k, \ \hat{\mu}_{+jk} = n_{+jk}。 \tag{8.24}$$

这些方程已经包含了对低阶项求导所对应的似然方程(习题 8.30)。对于模型(XZ,YZ),它的拟合值具有与观察数据相同的 XZ 和 YZ 的边际总计。

8.6.3 关于对数线性模型的 Birch 结果

对于模型(XZ,YZ),由式 8.23、式 8.24,以及表 8.12 可知,最小充分统计量是期望频数所对应的边际分布的最大似然估计。方程式 8.22 给出了关于任意对数线性模型的一般性结果。Birch(1963)证明,对数线性模型的似然方程就是使最小充分统计量等于它们的期望值。在式(4.29)和(4.44)中,有关泊松分布广义线性模型的理论也隐含着这一结果。因此,对数线性模型的拟合值是一种修匀的单元格计数,它们与观察到的单元格计数的某些边际分布相同,同时还满足模型所设定的关联和交互模式。

Birch 证明,存在一组唯一的拟合值既满足模型设定,又在最小充分统计量方面与数据相吻合。因此,如果我们找到这样的一个解,它一定就是最大似然解。具体地说,一个二维表格的独立性模型

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

具有最小充分统计量 $\{n_{i+}\}$ 和 $\{n_{+j}\}$。它的似然方程为:

$$对于所有 i 和 j, \quad \hat{\mu}_{i+} = n_{i+}, \quad \hat{\mu}_{+j} = n_{+j}。$$

相应的拟合值 $\{\hat{\mu}_{ij}=n_{i+}n_{+j}/n\}$ 既满足这些方程式,同时也满足模型本身。Birch 的结果意味着,它们就是最大似然估计值。

8.6.4 拟合值的直接计算法和迭代法

为了展示如何求解似然方程,我们继续以模型(XZ,YZ)为例。由式 8.9 可得,模型满足:

$$对于所有的 i, j 和 k, \quad \pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}。$$

就泊松分布样本而言,相应公式使用的是期望频数。令 $\pi_{ijk}=\mu_{ijk}/n$,则有 $\{\mu_{ijk}=\mu_{i+k}\mu_{+jk}/\mu_{++k}\}$。式 8.23 和式 8.24 似然方程设定最大似然估计满足 $\hat{\mu}_{i+k}=n_{i+k}$ 和 $\hat{\mu}_{+jk}=n_{+jk}$,进而 $\hat{\mu}_{++k}=n_{++k}$ 也成立。由于对包含参数的函数进行最大似然估计也就是对参数本身的最大似然估计,

$$\hat{\mu}_{ijk} = \frac{\hat{\mu}_{i+k}\hat{\mu}_{+jk}}{\hat{\mu}_{++k}} = \frac{n_{i+k}n_{+jk}}{n_{++k}}。$$

这个解既满足模型的要求,其最小充分统计量也与数据相一致。因此,它是唯一的最大似然解。

利用相似的推理过程，可以导出表 8.12 中几乎所有模型（其中一个除外）的 $\{\hat{\mu}_{ijk}\}$。表 8.13 给出了相应公式，并将 $\{\pi_{ijk}\}$ 表达为边际概率的形式。根据上面介绍的方法，这些表达式和似然方程决定了最大似然公式。

对于具有明确的关于 $\hat{\mu}_{ijk}$ 的公式的模型，我们称之为*直接*（*direct*）估计。许多对数线性模型不存在直接估计。这时，需要通过迭代法进行最大似然估计。在表 8.12 和表 8.13的模型中，唯一一个不存在直接估计的模型是（XY,XZ,YZ）。对此模型，尽管二维边际表给出了它的最小充分统计量，但 $\{\pi_{ijk}\}$ 无法直接用 $\{\pi_{ij+}\}$、$\{\pi_{i+k}\}$ 和 $\{\pi_{+jk}\}$ 来表示。对于包含所有二维关联项的非饱和模型，直接估计都不存在。在实际应用中，是否知道哪些模型具有直接估计并不重要，迭代法也适用于所有存在直接估计的模型。拟合对数线性模型的统计软件在*所有*情况下都使用这种迭代法。

表 8.13　关于三维表格的对数线性模型的拟合值

模　型[a]	概率形式	拟合值
(X,Y,Z)	$\pi_{ijk}=\pi_{i++}\pi_{+j+}\pi_{++k}$	$\hat{\mu}_{ijk}=\frac{n_{i++}n_{+j+}n_{++k}}{n^2}$
(XY,Z)	$\pi_{ijk}=\pi_{ij+}\pi_{++k}$	$\hat{\mu}_{ijk}=\frac{n_{ij+}n_{++k}}{n}$
(XY,XZ)	$\pi_{ijk}=\frac{\pi_{ij+}\pi_{i+k}}{\pi_{i++}}$	$\hat{\mu}_{ijk}=\frac{n_{ij+}n_{i+k}}{n_{i++}}$
(XY,XZ,YZ)	$\pi_{ijk}=\psi_{ij}\phi_{jk}\omega_{ik}$	迭代法　第 8.7 节
(XYZ)	无限定	$\hat{\mu}_{ijk}=n_{ijk}$

a 未列出的模型公式可根据对称性推出；例如，对于 (XZ,Y)，$\hat{\mu}_{ijk}=n_{i+k}n_{+j+}/n$。

8.6.5　卡方拟合优度检验

模型的拟合优度统计量将拟合的单元格计数与样本计数进行对比。在第 4.5.2 节中我们证明，对于泊松分布广义线性模型，当模型中包含截距项时，它的偏离度等于 G^2 统计量。在单元格的数目固定不变的情况下，当期望频数较大时，G^2 和 X^2 均近似服从卡方零分布。自由度等于零假设与备择假设的维度之差，也即，相当于一般性模型的参数数量减去被检验模型的参数数量。

我们通过模型 (X,Y,Z) 来加以说明，数据来自于概率为 $\{\pi_{ijk}\}$ 的多项分布样本。在一般性模型中，唯一的限定条件是 $\sum_i\sum_j\sum_k\pi_{ijk}=1$，所以它具有 $IJK-1$ 个参数。对于模型 (X,Y,Z)，$\{\pi_{ijk}=\pi_{i++}\pi_{+j+}\pi_{++k}\}$ 由 $I-1$ 个 $\{\pi_{i++}\}$（因为 $\sum_i\pi_{i++}=1$）、$J-1$ 个 $\{\pi_{+j+}\}$，以及 $K-1$ 个 $\{\pi_{++k}\}$ 所决定。因此，

$$\mathrm{df}=(IJK-1)-[(I-1)+(J-1)+(K-1)]=IJK-I-J-K+2。$$

同一自由度公式也适用于泊松分布样本。这时，一般性模型具有 IJK 个关于 $\{\mu_{ijk}\}$ 的参数。残差自由度等于表中所包括的单元格数量减去关于 $\{\mu_{ijk}\}$ 的泊松分布对数线性模型的参数数量。例如，模型 (X,Y,Z) 的残差自由度为 $\mathrm{df}=IJK-[1+(I-1)+(J-1)+(K-1)]$，反映了模型所包括的截距项 λ 以及诸如 $\lambda_I^X=\lambda_J^Y=\lambda_K^Z=0$ 的约束条件。这等于，为了得到某一特定模型，需要将饱和模型中线性独立的参数设定为 0 的数量。表 8.14给出了对三维对数线性模型进行检验的自由度公式。

表 8.14　关于三维表格的对数线性模型的残差自由度

模　型	自由度
(X,Y,Z)	$IJK-I-J-K+2$
(XY,Z)	$(K-1)(IJ-1)$
(XZ,Y)	$(J-1)(IK-1)$
(YZ,X)	$(I-1)(JK-1)$
(XY,YZ)	$J(I-1)(K-1)$
(XZ,YZ)	$K(I-1)(J-1)$
(XY,XZ)	$I(J-1)(K-1)$
(XY,XZ,YZ)	$(I-1)(J-1)(K-1)$
(XYZ)	0

8.6.6　最大似然参数估计的协方差矩阵

为了介绍最大似然参数估计值的大样本分布,我们回到一般表达式 $\log(\mu_i)=\sum_j x_{ij}\beta_j$,从中求得对数似然函数的导数为

$$\frac{\partial L(\boldsymbol{\mu})}{\partial\beta_j}=\sum_i n_i x_{ij}-\sum_i \mu_i x_{ij},\quad j=1,2,\cdots,p。$$

它的二阶偏导数的海塞矩阵的元素为

$$\frac{\partial^2 L(\boldsymbol{\mu})}{\partial\beta_j\partial\beta_k}=-\sum_i x_{ij}\frac{\partial\mu_i}{\partial\beta_k}$$

$$=-\sum_i x_{ij}\left\{\frac{\partial}{\partial\beta_k}\left[\exp\left(\sum_h x_{ih}\beta_h\right)\right]\right\}=-\sum_i x_{ij}x_{ik}\mu_i。$$

与 Logistic 回归模型相同,对数线性模型也是使用典型连结的广义线性模型,因而这个矩阵不依赖于观察到的数据。该矩阵的逆矩阵,即信息矩阵为

$$\mathfrak{I}=\mathbf{X}'\mathbf{diag}(\boldsymbol{\mu})\mathbf{X},$$

其中 $\mathbf{diag}(\boldsymbol{\mu})$ 主对角线上的元素为 $\boldsymbol{\mu}$。

在单元格数量固定的情况下,随着 $n\to\infty$,最大似然估计值 $\hat{\boldsymbol{\beta}}$ 渐近服从均值为 $\boldsymbol{\beta}$、协方差矩阵为 $\mathfrak{I}^{-1}$ 的正态分布。因此,对于泊松分布样本,渐近协方差矩阵为

$$\mathrm{cov}(\hat{\boldsymbol{\beta}})=[\mathbf{X}'\mathbf{diag}(\boldsymbol{\mu})\mathbf{X}]^{-1}。\tag{8.25}$$

代入最大似然拟合值并对主对角线上的元素求平方根,就得出了关于 $\hat{\boldsymbol{\beta}}$ 的标准误。如同在第 4.4.7 节所指出的,这也可以从广义线性模型的一般表达式(式 4.28)推导得出。

8.6.7　多项分布对数线性模型与泊松分布对数线性模型之间的联系

相似的渐近结果对于多项分布样本也成立。当 $\{Y_i,i=1,\cdots,N\}$ 是独立的泊松分布随机变量时,给定 $n=\sum_i Y_i$,$\{Y_i\}$ 的条件分布服从参数为 $\{\pi_i=\mu_i/(\sum_a\mu_a)\}$ 的多项分布。Birch(1963)证明,在多项分布样本情况下,对数线性模型参数的最大似然估计与独立的泊松分布样本时是相同的。他还证明,只要模型中包含了由抽样设计所固定的边际分布的项,独立的多项分布样本的估计结果也是相同的。具体而言,假定在 X 和 Z 的每个取值结合点,Y 来自于一个独立的多项分布样本。这时,$\{n_{i+k}\}$ 是固定的。模型必须包含 λ_{ik}^{XZ} 项,以使得拟合值满足 $\{\hat{\mu}_{i+k}=n_{i+k}\}$。

根据以下推理,不需要关于多项分布对数线性模型的单独的推断理论。将 $\{\mu_i\}$ 的泊

松分布对数线性模型表示为

$$\log \mu_i = \lambda + \mathbf{x}_i \boldsymbol{\beta},$$

其中,$(1,\mathbf{x}_i)$是模型矩阵 $\mathbf{X}$ 的第 i 行,$(\boldsymbol{\lambda},\boldsymbol{\beta}')'$是模型的参数向量。泊松分布对数似然函数可表示为

$$L = L(\lambda,\boldsymbol{\beta}) = \sum_i n_i \log \mu_i - \sum_i \mu_i$$
$$= \sum_i n_i(\lambda + \mathbf{x}_i\boldsymbol{\beta}) - \sum_i \exp(\lambda + \mathbf{x}_i\boldsymbol{\beta}) = n\lambda + \sum_i n_i \mathbf{x}_i \boldsymbol{\beta} - \tau,$$

其中,$\tau = \sum_i \mu_i = \sum_i \exp(\lambda + \mathbf{x}_i\boldsymbol{\beta})$。由于 $\log \tau = \lambda + \log[\sum_i \exp(\mathbf{x}_i\boldsymbol{\beta})]$,该对数似然函数具有以下形式:

$$L = L(\tau,\boldsymbol{\beta}) = \{\sum_i n_i \mathbf{x}_i \boldsymbol{\beta} - n\log[\sum_i \exp(\mathbf{x}_i\boldsymbol{\beta})]\} + (n\log \tau - \tau)。 \quad (8.26)$$

现在,$\pi_i = \mu_i/(\sum_a \mu_a) = \exp(\lambda + \mathbf{x}_i\boldsymbol{\beta})/[\sum_a \exp(\lambda + \mathbf{x}_a\boldsymbol{\beta})]$,并且分子和分母中的 $\exp(\lambda)$相互抵消。因此,式 8.26 等式右边的第一项(括号中的项)为 $\sum n_i \log \pi_i$,这就是以单元格总计数 n 为条件的多项分布对数似然函数。在无约束条件的情况下,$n = \sum_i n_i$ 服从期望值为 $\sum_i \mu_i = \tau$ 的泊松分布,所以式 8.26 中的第二项是关于 n 的泊松分布对数似然函数。由于 $\boldsymbol{\beta}$ 只出现在第一项中,泊松分布对数似然函数 $L(\lambda,\boldsymbol{\beta})$的最大似然估计值 $\boldsymbol{\beta}$ 及其协方差矩阵与多项分布对数似然函数的是完全一致的。由于泊松分布样本中样本规模是随机的,泊松分布对数线性模型比多项分布对数线性模型多一个参数(即 λ)。有关细节,参见:Birch(1963)、Lang(1996c)、McCullagh and Nelder(1989:211)、Palmgren(1981)。

我们将在第 14.4.1 节给出,对于多项分布样本,它的对数线性模型参数估计值的估计协方差矩阵为

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = \left\{\mathbf{X}'\left[\mathbf{diag}(\hat{\boldsymbol{\mu}}) - \frac{\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}'}{n}\right]\mathbf{X}\right\}^{-1}。 \quad (8.27)$$

泊松分布模型中的截距项 λ 与此无关,并且多项分布模型中的 $\mathbf{X}$ 对应于在泊松分布模型的 $\mathbf{X}$ 中删除包含 λ 的那一列。

相似的推理也适用于存在多个独立的多项分布样本的情况。对数似然函数的每一项都是来自于不同样本的相应项的加总,不过,泊松分布对数似然函数可再次分解为两个部分。一部分是关于独立样本规模的泊松分布对数似然函数,另一部分是独立的多项分布对数似然函数的加总。Palmgren(1981)证明,以观察到的解释变量的边际总计为条件,与结果变量有关的参数估计值的渐近协方差与泊松分布样本的情况相同。对于单个的多项分布样本,Palmgren 的结果意味着式 8.27 与删除掉包括 λ 的行和列后的式 8.25 完全一致。Birch(1963)以及 Goodman(1970)给出了相应的结果。Lang(1996c)就多项分布模型和泊松分布模型之间的关系进行了精彩的讨论。他的结果表明,给定协变量的取值水平,两模型估计的对数均值的任何线性函数的渐近方差都相同。

8.6.8 概率估计值的分布

对于多项分布样本,单元格概率的最大似然估计值为 $\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\mu}}/n$。接下来,我们给出 $\mathrm{cov}(\hat{\boldsymbol{\pi}})$的渐近值。Lang(1996c)介绍了在泊松分布样本中 $\hat{\boldsymbol{\mu}}$ 的渐近协方差矩阵以及它与 $\mathrm{cov}(\hat{\boldsymbol{\pi}})$的关系。

饱和模型具有 $\hat{\boldsymbol{\pi}} = \mathbf{p}$,即估计概率等于样本比例。在多项分布样本中,由式 3.7 和式

3.8 可得，相应的协方差矩阵为

$$\mathrm{cov}(\mathbf{p}) = \frac{[\mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}']}{n}。 \tag{8.28}$$

对于一个具有 J 个类别的结果变量的 I 个独立的多项分布样本，$\boldsymbol{\pi}$ 和 $\mathbf{p}$ 由 I 组比例构成，每组具有 $J-1$ 个非冗余元素。这时，$\mathrm{cov}(\mathbf{p})$ 是一个分块对角矩阵（block diagonal matrix）。每个独立样本对应着一块形式为式 8.28 的 $(J-1)\times(J-1)$ 矩阵，每块矩阵中主对角线以外的元素为零。

现在假定存在一个非饱和模型，$\hat{\boldsymbol{\pi}}$ 渐近服从关于 $\boldsymbol{\pi}$ 的正态分布，这一点我们将在第 14.2.2 和第 14.4.1 节中利用 δ 方法加以证明，所估计的协方差矩阵等于

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\pi}}) = \frac{\{\widehat{\mathrm{cov}}(\mathbf{p})\mathbf{X}[\mathbf{X}'\widehat{\mathrm{cov}}(\mathbf{p})\mathbf{X}]^{-1}\mathbf{X}'\widehat{\mathrm{cov}}(\mathbf{p})\}}{n}。$$

在单个多项分布样本的情况下，该表达式等于

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\pi}}) = \frac{\{[\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}']\mathbf{X}[\mathbf{X}'(\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}')\mathbf{X}]^{-1}\mathbf{X}'[\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}']\}}{n}。$$

当表格包括很多单元格时，往往会出现其中某个单元格的样本比例为 0 的情况。在这种情况下，普通的标准误为 0，这一结果难以令人满意。模型拟合的一个优点在于，它通常会给出正的拟合概率和标准误。

8.6.9 最大似然估计的唯一性

当存在所有的 $\{n_i > 0\}$ 时，最大似然估计值存在并且唯一。为简便起见，我们使用泊松分布样本对此进行证明。假定模型的参数设定确保 $\mathbf{X}$ 是满秩矩阵，Birch（1963）证明了似然方程是有解的。他指出，泊松分布对数似然函数的核函数

$$L(\boldsymbol{\mu}) = \sum_i (n_i \log \mu_i - \mu_i)$$

的每一项随着 $\log(\mu_i) \to \pm\infty$ 而收敛于 $-\infty$，因此，上述对数似然函数是有边界的，并且在模型参数取有限值时达到最大值。该函数在这个最大值上是稳定的，因为它具有连续的一阶偏导数。

Birch 显示，似然方程具有唯一解，并且似然函数在该解处达到最大值。通过给出关于 $\{-\partial^2 L/\partial\beta_h\partial\beta_j\}$ 的矩阵（即，信息矩阵 $\mathbf{X}'\mathbf{diag}(\boldsymbol{\mu})\mathbf{X}$）是非奇异、非负定的，因而必然是正定的，他证明了上述结果。矩阵的非奇异性来自于 $\mathbf{X}$ 是满秩矩阵，并且对角矩阵具有正的元素 $\{\mu_i\}$。任意二次项形式的 $\mathbf{c}'\mathbf{X}'\mathbf{diag}(\boldsymbol{\mu})\mathbf{X}\mathbf{c}$ 等于 $\sum_i[\sqrt{\mu_i}(\sum_j x_{ij}c_j)]^2 \geqslant 0$，因而该矩阵同时也是非负定的。

8.7 对数线性模型的拟合：迭代法及其应用*

当一个对数线性模型不存在直接估计时，可以利用迭代法来求解似然方程，比如 Newton-Raphson 算法。在本节中，我们还将介绍一种较为简便，但也具有较大局限性的方法，*迭代成比例拟合法*（*iterative proportional fitting*）。

8.7.1 Newton-Raphson 算法

在第 4.6.1 节中，我们介绍了 Newton-Raphson 算法。沿用当时的表达符号，我们令 $L(\boldsymbol{\beta})$ 表示泊松分布对数线性模型的对数似然函数。

由式 8.21 可得，

$$L(\boldsymbol{\beta}) = \sum_i n_i(\sum_h x_{ih}\beta_h) - \sum_i \exp(\sum_h x_{ih}\beta_h)。$$

那么，

$$u_j = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i n_i x_{ij} - \sum_i \mu_i x_{ij},$$

$$h_{jk} = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = -\sum_i \mu_i x_{ij} x_{ik},$$

因而，

$$u_j^{(t)} = \sum_i (n_i - \mu_i^{(t)}) x_{ij} \quad 并且 \quad h_{jk}^{(t)} = -\sum_i \mu_i^{(t)} x_{ij} x_{ik}。$$

对 $\hat{\boldsymbol{\mu}}$ 的第 t 次近似 $\boldsymbol{\mu}^{(t)}$ 是由 $\boldsymbol{\beta}^{(t)}$ 推导而来的，即有 $\boldsymbol{\mu}^{(t)} = \exp(\mathbf{X}\boldsymbol{\beta}^{(t)})$。它通过式 4.39 生成下一个值 $\boldsymbol{\beta}^{(t+1)}$，在这里

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + [\mathbf{X}'\mathbf{diag}(\boldsymbol{\mu}^{(t)})\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{n} - \boldsymbol{\mu}^{(t)})。$$

再由此得出 $\boldsymbol{\mu}^{(t+1)}$，以此类推。

另外，$\boldsymbol{\beta}^{(t+1)}$ 也可以表示为

$$\boldsymbol{\beta}^{(t+1)} = -(\mathbf{H}^{(t)})^{-1}\mathbf{r}^{(t)}, \tag{8.29}$$

其中 $r_j^{(t)} = \sum \mu_i^{(t)} x_{ij}[\log \mu_i^{(t)} + (n_i - \mu_i^{(t)})/\mu_i^{(t)}]$。括号中的表达式是 $\log n_i$ 在 $\log \mu_i^{(t)}$ 处的泰勒级数展开的第一项。

拟合的步骤首先是令所有的 $\mu_i^{(0)} = n_i$。如果存在任何 $n_i = 0$，可以考虑用 $\mu_i^{(0)} = n_i + \frac{1}{2}$ 来进行调整。接着，利用式 8.29 求出 $\boldsymbol{\beta}^{(1)}$，然后按照上面的描述进行对 $t>0$ 的迭代过程。对于对数线性模型，$L(\boldsymbol{\beta})$ 是一个凹函数，并且随着 t 的增加，$\boldsymbol{\mu}^{(t)}$ 和 $\boldsymbol{\beta}^{(t)}$ 通常会迅速收敛于它们的最大似然估计 $\hat{\boldsymbol{\mu}}$ 和 $\hat{\boldsymbol{\beta}}$。矩阵 $\mathbf{H}^{(t)}$ 收敛于 $\hat{\mathbf{H}} = -\mathbf{X}'\mathbf{diag}(\hat{\boldsymbol{\mu}})\mathbf{X}$。按照式 8.25，迭代方法同时给出关于 $\hat{\boldsymbol{\beta}}$ 的大样本协方差矩阵的估计值 $-\hat{\mathbf{H}}^{-1}$。

如同我们在第 4.6.3 节关于广义线性模型的讨论，式 8.29 具有迭代再加权最小二乘法的形式：

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}'\hat{\mathbf{V}}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}_t^{-1}\mathbf{z}^{(t)}。$$

这里，$\mathbf{z}^{(t)}$ 的元素为 $n_i = \log \mu_i^{(t)} + (n_i - \mu_i^{(t)})/\mu_i^{(t)}$，并且 $\hat{\mathbf{V}}_t = [\mathbf{diag}(\boldsymbol{\mu}^{(t)})]^{-1}$。因此，$\boldsymbol{\beta}^{(t+1)}$ 是模型

$$\mathbf{z}^{(t)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

的加权最小二乘解，其中 $\{\varepsilon_i\}$ 与方差 $\{1/\mu_i^{(t)}\}$ 之间不存在相关关系。在 $\{\mu_i^{(0)} = n_i\}$ 的情况下，$\boldsymbol{\beta}^{(1)}$ 是对模型 $\log(\mathbf{n}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 的加权最小二乘估计。

8.7.2 迭代成比例拟合

迭代成比例拟合算法（*iterative proportional fitting algorithm*, *IPF*）是用来计算对数线性模型的 $\{\hat{\mu}_i\}$ 的一种简便方法。它由 Deming 和 Stephen（1940）提出，包括以下步骤：

1. 以一个比所要拟合的模型更简单的模型的 $\{\mu_i^{(0)}\}$ 为初值。例如，完全可以设 $\{\mu_i^{(0)} \equiv 1.0\}$。
2. 对 $\{\mu_i^{(0)}\}$ 乘以适当的因子进行逐步调整，使其与一组最小充分统计量中的每个边际表相一致。
3. 继续上述过程，直到充分统计量与它们的拟合值之差足够接近于零。

我们以模型(XY,XZ,YZ)为例来加以说明。该模型的最小充分统计量为$\{n_{ij+}\}$、$\{n_{i+k}\}$,以及$\{n_{+jk}\}$。初始估计必须使这个模型成立。IPF 算法的第一个循环包括以下三步:

$$\mu_{ijk}^{(1)}=\mu_{ijk}^{(0)}\frac{n_{ij+}}{\mu_{ij+}^{(0)}},\ \mu_{ijk}^{(2)}=\mu_{ijk}^{(1)}\frac{n_{i+k}}{\mu_{i+k}^{(1)}},\ \mu_{ijk}^{(3)}=\mu_{ijk}^{(2)}\frac{n_{+jk}}{\mu_{+jk}^{(2)}}。$$

对于所有的 i 和 j,在第一个表达式的两端对 k 求和,有 $\mu_{ij+}^{(1)}=n_{ij+}$。第一步后,在 XY 的边际表中,观察值与拟合值相同。在第二步后,所有的 $\mu_{i+k}^{(2)}=n_{i+k}$,但 XY 边际表不再相等。第三步后,所有的 $\mu_{+jk}^{(3)}=n_{+jk}$,但 XY 和 XZ 边际表不再相等。然后,通过 $\mu_{ijk}^{(4)}=\mu_{ijk}^{(3)}(n_{ij+}/\mu_{ij+}^{(3)})$ 重新令 XY 边际表相合,开始一个新的循环,逐步类推。

在每一步中,更新后的估计值继续满足模型的条件。例如,第一步针对 Z 的不同取值 k 使用了同一个调整因子($n_{ij+}/\mu_{ij+}^{(0)}$)。因而,在 Z 的不同取值上,XZ 的发生比之比的比值等于 1,这样,每一步都保证了同质性关联的模式。

随着循环的进行,比较单元格计数及其拟合值的 G^2 统计量单调递减,并且整个过程一定收敛(Fienberg,1970a;Haberman,1974a)。通过 IPF 算法会得出最大似然估计值,因为它所生成的一系列拟合值收敛于一个解,这个解既满足模型设定又与充分统计量相一致。按照 Birch 的结果(第 8.6.3 节),只有一个这样的解,它就是最大似然估计值。

即便模型存在直接估计,也可以使用 IPF 算法。在这种情况下,IPF 往往会在一个循环内给出最大似然估计(Haberman,1974a,p. 197)。我们通过独立性模型来加以演示。它的最小充分统计量为$\{n_{i+}\}$和$\{n_{+j}\}$。利用$\{\mu_{ij}^{(0)}\equiv 1.0\}$,第一个循环给出

$$\mu_{ij}^{(1)}=\mu_{ij}^{(0)}\frac{n_{i+}}{\mu_{i+}^{(0)}}=\frac{n_{i+}}{J},$$

$$\mu_{ij}^{(2)}=\mu_{ij}^{(1)}\frac{n_{+j}}{\mu_{+j}^{(1)}}=\frac{n_{i+}n_{+j}}{n}。$$

这时对于所有 $t>2$,IPF 算法都给出 $\hat{\mu}_{ij}^{(t)}=n_{i+}n_{+j}/n$

8.7.3 迭代法的比较

IPF 算法很简单,而且便于操作,甚至当似然函数出现问题——比如拟合计数为零以及估计值发生在参数空间的边界时,它仍然收敛于最大似然拟合。相对而言,Newton-Raphson 算法更加复杂,它需要在每一步都求解一个方程组。当模型包括很高的维度时——例如,当列联表和参数空间非常庞大时,Newton-Raphson 算法有点不切实际。

但是,IPF 算法也有它的缺点。它主要适用于那些似然方程为令边际表中的观察计数等于拟合计数的模型。相反,Newton-Raphson 算法是一种非常一般性的方法,它可以用来求解更复杂的似然方程。与 Newton-Raphson 算法相比,IPF 算法有时收敛的速度较慢。与 Newton-Raphson 算法不同的是,IPF 本身并不给出模型的参数估计以及相应的协方差矩阵估计。当然,利用 IPF 算法的拟合值可以计算这些信息。模型的参数估计可以根据$\{\log\hat{\mu}_i\}$回推得出(参见习题 8.16 和 8.17),将拟合值代入式 8.25 可以得到$\mathrm{cov}(\hat{\boldsymbol{\beta}})$。

由于 Newton-Raphson 算法的适用范围非常广泛,并且它还给出标准误估计,所以在大多数拟合对数线性模型的软件中,它是常规的拟合方法。目前,IPF 算法越来越普遍地被视为一种过时的方法。不过,在一些应用中,利用 IPF 算法会使分析的意义更加明确,比如下面所要讲到的例子。

8.7.4 列联表的标准化

根据美国民意研究中心(National Opinion Research Center)进行的综合社会调查(General Social Survey),表 8.15 给出了教育程度与对待合法性人工流产的态度之间的关系。为了使关联模式更清楚,Smith(1976)对该表进行了标准化,以使得在维持样本发生比之比结构不变的同时,令所有行和列的边际总计都等于 100。

表 8.15 对表格边际进行标准化后按照教育程度划分的关于人工流产的态度

教育程度	关于合法流产的态度			
	反对	中间立场	支持	小计
高中以下	209	101	237	
	(49.4)	(32.0)	(18.6)	(100)
高中	151	126	426	
	(32.8)	(36.6)	(30.6)	(100)
高中以上	16	21	138	
	(17.8)	(31.3)	(50.9)	(100)
小计	(100)	(100)	(100)	

来源:Smith(1976)。

实现使表格边际总计等于 100 的标准化的 IPF 算法为

$$\mu_{ij}^{(0)} = n_{ij},$$

接着对于 $t=1,3,5,\cdots$,

$$\mu_{ij}^{(t)} = \mu_{ij}^{(t-1)} \frac{100}{\mu_{i+}^{(t-1)}}, \quad \mu_{ij}^{(t+1)} = \mu_{ij}^{(t)} \frac{100}{\mu_{+j}^{(t)}}。$$

在每个奇数步完成后,所有的行总计等于 100。在每个偶数步完成后,所有的列总计等于 100。发生比之比在每一步中都保持不变,因为每个给定行(列)中的所有计数都乘以了一个相同的常数。

IPF 算法收敛于表 8.15 中括号里的值。在这个标准化的表格中,关联模式更加清楚。在主对角线上出现了一道脊,其中具有较高教育程度的人更倾向于支持人工流产。其他的计数均匀地落在主对角线的两侧。

在比较具有不同边际结构的表格时,对表格进行标准化很有帮助。Mosteller(1968)对比了英国和丹麦的代际职业流动表。Yule 比较了三家医院里天花病人接种疫苗与康复的情况。表格标准化的一种现代应用是对样本数据进行调整,以使它的边际分布与普查结果相一致。

对表格进行标准化的过程被称为*标准化*(*raking*)表格。Imrey 等(1981)以及 Little 和 Wu (1991)推导出了标准化样本比例的渐近协方差矩阵。对于满足$\{\mu_{ij}=E(n_{ij})\}$的样本计数$\{n_{ij}\}$,令$\{E_{ij}\}$表示标准化后表格的期望频数,$\{\hat{E}_{ij}\}$为标准化表格的拟合值。标准化过程相当于拟合模型

$$\log(E_{ij}/\mu_{ij}) = \lambda + \lambda_i^E + \lambda_j^A。$$

也即,保持发生比之比不变意味着,关于$\{E_{ij}/\mu_{ij}\}$和$\{\hat{E}_{ij}/n_{ij}\}$的二维表格满足独立性。

进行标准化后,表格的拟合值$\{\hat{E}_{ij}\}$满足

$$\log \hat{E}_{ij} - \log n_{ij} = \hat{\lambda} + \hat{\lambda}_i^E + \hat{\lambda}_j^A。$$

对该拟合的对数连结的调整项，$-\log n_{ij}$，被称为*抵消项*（*offset*）。这个拟合相当于将 $\log n_{ij}$作为等式右边的一个预测项，并将其系数强制设定为 1.0。标准的广义线性模型（GLM）软件可以拟合包含抵消项的模型。在对表格进行标准化时，可以输入满足独立性并且具有想要的边际分布的样本数据假值（pseudo-values），将 $\log n_{ij}$视为抵消项（有关的 SAS 程序，参见附表 A.14）。在第 9.7.1 节中，我们将进一步讨论模型抵消项的有关应用。

注 解

第 8.2 节：关于三维表的独立性和包括交互效应的对数线性模型

8.1 Roy 和 Mitra（1956）讨论了有关三维表格中独立性的种类以及相应的大样本检验。Birch（1963）关于对数线性模型的最大似然估计的研究属于在 1960 年代大量有关对数线性模型研究的一部分，其发展在很大程度上归功于 L. A. Goodman（参见第 16.4 节）。Haberman（1974a）为对数线性模型作出了重大的理论贡献。

第 8.3 节：对数线性模型的统计推断

8.2 Goodman（1970，1971b）、Haberman（1974a，chap. 5）、Lauritzen（1996）、Sundberg（1975），以及 Whittaker（1990，sec. 12.4）讨论了存在直接的最大似然估计的对数线性模型，并按照了有关独立性、条件独立性或等概率性进行了解释。这些模型被称为*可分解*（*decomposable*）模型，因为它们的期望频数可以分解为期望的边际充分统计量的乘积和比率。Haberman 证明了对数线性模型存在直接估计的条件。Baglivo 等（1992）、Forster 等（1996），以及 Morgan 和 Blumensein（1991）讨论了精确推断的方法。

8.3 有关允许错误设定误差（misclassification error）的方法，参见：Kuha and Skinner（1997）、Kuha et al.（1998），以及他们所引述的文献。关于缺失数据的处理，参见：Little（1998）、Schafer（1997，chap. 8），以及其中的文献。

第 8.7 节：对数线性模型的拟合：迭代法及其应用

8.4 Deming（Deming，1964，chap. Ⅶ）描述了 Deming 和 Stephan 关于 IPF 算法的早期研究。Darroch（1962）利用 IPF 算法得出了关于列联表的最大似然估计。Bishop 等（1975）、Fienberg（1970a），以及 Speed（1998）介绍了有关 IPF 算法的其他应用。Darroch 和 Ratcliff（1972）将 IPF 算法扩展到充分统计量比边际表更为复杂的模型。

8.5 有关标准化表格的进一步讨论，参见：Bishop et al.（1975：76-102）、Fleiss（1981，chap. 14）、Haberman（1979，chap. 9）、Hoem（1987）、Little and Wu（1991）。

习 题

应用部分

8.1 美国民意研究中心（National Opinion Research Center）在 1988 年综合社会调查（General Social Survey）中问被访者："你支不支持下列关于艾滋病的举措？（1）由政府支付艾滋病患者的所有医疗费用；（2）发起一个政府信息项目以提倡安全性行为，比如使用安全套。"表 8.16 给出了分性别（G）的关于医疗费用（H）和信息项目（I）的回答结果。

a. 拟合对数线性模型（GH，GI），（GH，HI），（GI，HI），以及（GH，GI，HI）。指出不包括 HI 项的模型拟合结果很差。

b. 利用模型(GH,GI,HI),给出关于 GH 的条件发生比之比的95%的沃尔德置信区间等于(0.55,1.10),关于 GI 的条件发生比之比的置信区间为(0.99,2.55)。对结果加以解释。关于这些观点,有没有可能不存在性别差异?

表 8.16　习题 8.1 的数据

性别	对信息项目的态度	对医疗支出的态度	
		支持	反对
男性	支持	76	160
	反对	6	25
女性	支持	114	181
	反对	11	48

来源:1988 年综合社会调查(General Social Survey),美国民意研究中心(National Opinion Research Center)。

8.2 参见表 8.17,该数据取自 1991 年综合社会调查(General Social Survey)。白人被访者被问到:(B)“你愿意开校车送一个学区的黑人和白人学生去另一个学区吗?”,(P)“如果你所在的党提名一个黑人作为总统候选人并且他能胜任的话,你会投票给他吗?”,(D)“在过去的几年里,你们家有没有人带黑人朋友回家吃饭?”。每个问题的选项都为(是,否,不知道)。拟合模型(BD,BP,DP)。

a. 利用回答为“是”和“否”的数据,估计关于每对变量的条件发生比之比。解释结果。

b. 分析模型的拟合优度。解释结果。

c. 利用沃尔德或似然比置信区间,对 BP 的条件关联进行统计推断并加以检验。解释结果。

表 8.17　习题 8.2 的数据[a]

总统候选人	校车	带朋友回家		
		1	2	3
1	1	41	65	0
	2	71	157	1
	3	1	17	0
2	1	2	5	0
	2	3	44	0
	3	1	0	0
3	1	0	3	1
	2	0	10	0
	3	0	0	1

a 1,是;2,否;3,不知道。

来源:1991 年综合社会调查(General Social Survey),美国民意研究中心(National Opinion Research Center)。

8.3 参考第 8.3.2 节,说明为什么以变量不同类别的参数之和等于零作为限定条件的统计软件会报告 $\hat{\lambda}_{11}^{AC}=\hat{\lambda}_{22}^{AC}=0.514$ 以及 $\hat{\lambda}_{12}^{AC}=\hat{\lambda}_{21}^{AC}=-0.514$,每一项的 SE $=0.044$。

8.4 参见表 2.6。令 $D=$ 被告人种族,$V=$ 受害人种族,以及 $P=$ 死刑判决结果。拟合对数线性模型(DV,DP,PV)。

a. 利用模型的拟合值,在 V 的每个取值上,估计关于 D 和 P 的发生比之比并加以解释。指出它们是否具有共同发生比之比的特性。

b. 计算关于 D 和 P 的边际发生比之比:(i)利用拟合值,(ii)利用样本数据。为什么它们相等?构建关于(a)部分的发生比之比。说明为什么会出现辛普森悖论。

c. 拟合相应的 Logit 模型,以 P 作为结果变量。给出参数估计与拟合统计量之间的对应关系。

d. 是否存在一个拟合得较好的更简单模型?加以解释,并指出 Logit 模型与对数线性模型之间的联系。

8.5 表 8.18 所示为 1988 年佛罗里达州发生的车祸的有关统计数据。

表 8.18 习题 8.5 的数据

安全装备的使用	是否被弹出	受伤状况	
		非致命伤	致命伤
安全带	是	1 105	14
	否	411 111	483
无	是	4 624	497
	否	157 342	1 008

来源:佛罗里达州高速公路安全与机动车管理局。

a. 找出一个能很好描述该数据的对数线性模型,解释其中的关联。

b. 将是否死亡作为结果变量,拟合一个等价的 Logit 模型。解释模型中的有关效应。

c. 由于 n 很大,除非模型拟合得非常好,拟合优度统计量都会很大。计算(a)部分中模型的差异指数,并加以解释。

8.6 参考表 8.19。调查对象被问及他们关于政府在环境(E)、健康(H)、对大城市的援助(C),以及法律执行(L)方面花费的观点。

表 8.19 习题 8.6 的数据[a]

环境	健康	法律执行	大城市援助								
			1			2			3		
			1	2	3	1	2	3	1	2	3
1	1		62	17	5	90	42	3	74	31	11
	2		11	7	0	22	18	1	19	14	3
	3		2	3	1	2	0	1	1	3	1
2	1		11	3	0	21	13	2	20	8	3
	2		1	4	0	6	9	0	6	5	2
	3		1	0	1	2	1	1	4	3	1
3	1		3	0	0	2	1	0	9	2	1
	2		1	0	0	2	1	0	4	2	0
	3		1	0	0	0	0	0	1	2	3

a 1,太少了;2,差不多;3,太多了。

来源:1989 年综合社会调查(General Social Survey),美国民意研究中心(National Opinion Research Center)。

a. 表 8.20 给出了一些模型的结果,包括关于同质性关联模型的二维关联项的估计值。检查模型的拟合情况,并加以解释。

b. 对每个变量的第 3 个类别的所有参数估计都等于 0。利用每对变量的"太多了"和"太少了"两个类别,报告估计的条件发生比之比。对这些关联进行总结。基于这些结果,你会考虑在模型中删除掉哪些项?为什么?

c. 表 8.21 报告了当每行和每列的参数加总等于零以及当第一行和第一列的参数为零时的$\{\hat{\lambda}_{eh}^{EH}\}$。在这些情况下，指出如何估计关于“太多了”和“太少了”之间的 *EH* 条件发生比之比。将结果与(b)部分进行比较。构建关于该发生比之比的置信区间。加以解释。

表 8.20　对表 8.19 数据拟合模型的输出结果

Criteria For Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	48	31.669 5	0.659 8
Pearson Chi-Square	48	26.522 4	0.552 6
Log Likelihood		1 284.940 4	

parameter			DF	Estimate	Standard Error	Wald 95% Confidence	Limits	Chi-Square
e×h	1	1	1	2.142 5	0.556 6	1.051 5	3.233 5	14.81
e×h	1	2	1	1.422 1	0.603 4	0.239 4	2.604 9	5.55
e×h	2	1	1	0.729 4	0.566 7	−0.381 3	1.840 2	1.66
e×h	2	2	1	0.318 3	0.621 1	−0.899 1	1.535 6	0.26
e×l	1	1	1	−0.132 8	0.637 8	−1.382 9	1.117 2	0.04
e×l	1	2	1	0.373 9	0.697 5	−0.993 1	1.741 0	0.29
e×l	2	1	1	−0.263 0	0.679 6	−1.594 9	1.068 9	0.15
e×l	2	2	1	0.425 0	0.736 1	−1.017 8	1.867 8	0.33
e×c	1	1	1	1.200 0	0.517 7	0.185 4	2.214 7	5.37
e×c	1	2	1	1.389 6	0.477 4	0.454 0	2.325 3	8.47
e×c	2	1	1	0.691 7	0.560 5	−0.406 8	1.790 2	1.52
e×c	2	2	1	1.376 7	0.502 4	0.392 1	2.361 4	7.51
h×c	1	1	1	−0.186 5	0.454 7	−1.077 7	0.704 8	0.17
h×c	1	2	1	0.746 4	0.480 8	−0.195 9	1.688 6	2.41
h×c	2	1	1	−0.467 5	0.497 8	−1.443 1	0.508 1	0.88
h×c	2	2	1	0.729 3	0.502 3	−0.255 3	1.713 8	2.11
h×l	1	1	1	1.874 1	0.507 9	0.878 6	2.869 6	13.61
h×l	1	2	1	1.036 6	0.526 2	0.005 2	2.068 0	3.88
h×l	2	1	1	1.937 1	0.622 6	0.716 8	3.157 4	9.68
h×l	2	2	1	1.823 0	0.635 5	0.577 5	3.068 6	8.23
c×l	1	1	1	0.873 5	0.460 4	−0.028 9	1.776 0	3.60
c×l	1	2	1	0.570 7	0.486 3	−0.382 4	1.523 9	1.38
c×l	2	1	1	1.079 3	0.432 6	0.231 4	1.927 1	6.23
c×l	2	2	1	1.205 8	0.446 2	0.331 2	2.080 4	7.30

8.7　参见关于表 8.8 的对数线性模型。

a. 说明为什么在表 8.10 中模型(*GI*, *GL*, *GS*, *IL*, *IS*, *LS*)所拟合的发生比之比显示，最可能发生死亡的车祸情况是在农村地区不系安全带的女性。

b. 拟合模型(*GLS*, *GI*, *IL*, *IS*)。利用模型的参数估计，表明拟合的 *IS* 条件发生比之比等于 0.44。对于受伤的每个类别，给出所估计的 *LS* 条件发生比之比对女性为 1.17，而对男性为 1.03。你怎样来求出这些结果？

表 8.21 习题 8.6 的参数估计

E	限定条件为参数之和等于零			限定条件为第一类别参数等于零		
	H			H		
	1	2	3	1	2	3
1	0.509	0.166	-0.676	0	0	0
2	-0.065	-0.099	0.163	0	0.309	1.413
3	-0.445	-0.068	0.513	0	0.720	2.142

8.8 考虑关于表 8.8 的以下二阶段(two-stage)模型。第一阶段是 *GLS* 三维表格中以 *S* 作为结果变量的 Logit 模型。第二阶段是在四维表格中,将这三个变量作为 *I* 的预测变量的 Logit 模型。说明为什么这个复合模型是合理的,拟合这些模型,并对结果加以解释。

8.9 参考习题 5.24 中的 Logit 模型。令 *A* = 对人工流产的态度。

a. 用符号表示出与这个 Logit 模型等价的对数线性模型。

b. 哪一个 Logit 模型等价于对数线性模型(*AR*,*AP*,*GRP*)?

c. 给出在以下情况下等价的对数线性模型和 Logit 模型:(i)*A* 联合独立于 *G*,*R*,和 *P*;(ii)*R* 对 *A* 具有主效应,但给定 *R* 后,*A* 条件独立于 *G* 和 *P*;(iii)*P* 和 *R* 对 *A* 具有交互效应,并且 *G* 对 *A* 具有主效应。

8.10 分析多维列联表时,在什么情况下 Logit 模型比对数线性模型更合适?什么情况下对数线性模型比 Logit 模型更合适?

8.11 利用统计软件,完成在本章中所描述的关于学生调查数据的分析(表 8.3)。

8.12 对表 10.6 进行标准化。描述其中的迁移模式。

8.13 本书的网页(*www.stat.ufl.edu/~aa/cda/cda.html*)包括一个关于参加宗教活动的频率、政治观念、对面向青少年开展节育服务的态度,以及对婚前性行为的态度的 2×3×2×2 表格。利用对数线性模型分析这些数据。

理论与方法

8.14 假定$\{\mu_{ij}=n\pi_{ij}\}$满足独立性模型式 8.1。

a. 证明 $\lambda_a^Y-\lambda_b^Y=\log(\pi_{+a}/\pi_{+b})$。

b. 证明{所有的 $\lambda_j^Y=0$}等价于,对于所有 j,都有 $\pi_{+j}=1/J$。

8.15 参考独立性模型,$\mu_{ij}=\mu\alpha_i\beta_j$。对于相应的对数线性模型式 8.1:

a. 证明可以通过以下条件来限定 $\sum\lambda_i^X=\sum\lambda_j^Y=0$:

$$\lambda_i^X=\log\alpha_i-(\sum_h\log\alpha_h)/I,\lambda_j^Y=\log\beta_j-(\sum_h\log\beta_h)/J,$$

$$\lambda=\log\mu+\frac{(\sum_h\log\alpha_h)}{I}+\frac{(\sum_h\log\beta_h)}{J}$$

b. 证明可以通过定义 $\lambda_i^X=\log\alpha_i-\log\alpha_1$ 和 $\lambda_j^Y=\log\beta_j-\log\beta_1$ 来限定 $\lambda_1^X=\lambda_1^Y=0$。这时,$\lambda$ 等于什么?

8.16 对于一个 $I\times J$ 表格,令 $\eta_{ij}=\log\mu_{ij}$,并令“.”下标表示该指标的均值(如 $\eta_{i.}=\sum_j\eta_{ij}/J$)。那么,令 $\lambda=\eta_{..}$,$\lambda_i^X=\eta_{i.}-\eta_{..}$,$\lambda_j^Y=\eta_{.j}-\eta_{..}$,以及 $\lambda_{ij}^{XY}=\eta_{ij}-\eta_{i.}-\eta_{.j}+\eta_{..}$。

a. 证明 $\log\mu_{ij}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_{ij}^{XY}$。因此,任意一组正的$\{\mu_{ij}\}$都满足饱和模型。

b. 证明 $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$。

c. 对于 2×2 表格，证明 $\log\theta = 4\lambda_{11}^{XY}$。

d. 对于 $2\times J$ 表格，证明 $\lambda_{11}^{XY} = (\sum_j \log\alpha_j)/2J$，其中 $\alpha_j = \mu_{11}\mu_{2j}/\mu_{21}\mu_{1j}, j = 2,\cdots,J$。

e. 不同的限定条件对应着不同的发生比之比的公式。令 $\lambda = \eta_{11}, \lambda_i^X = \eta_{i1} - \eta_{11}$，$\lambda_j^Y = \eta_{1j} - \eta_{11}$，以及 $\lambda_{ij}^{XY} = \eta_{ij} - \eta_{i1} - \eta_{1j} + \eta_{11}$。这时，证明对于所有的 i 和 j，当 $\lambda_1^X = \lambda_1^Y = \lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0$ 时饱和模型成立，并且 $\lambda_{ij}^{XY} = \log(\mu_{11}\mu_{ij}/\mu_{1j}\mu_{i1})$。

8.17 假定所有的 $\mu_{ijk} > 0$。令 $\eta_{ijk} = \log\mu_{ijk}$，并考虑模型的限定条件是设参数之和为零。

a. 对于一般的对数线性模型式 8.12，按照习题 8.16 的方式定义参数（如 $\lambda_{ij}^{XY} = \eta_{ij.} - \eta_{i..} - \eta_{.j.} + \eta_{...}$）。

b. 对于一个 $2\times2\times2$ 表格所对应的模型（XY,XZ,YZ），证明 $\lambda_{11}^{XY} = \frac{1}{4}\log\theta_{11(k)}$。

c. 对于一个 $2\times2\times2$ 表格所对应的模型（XYZ），证明

$$\lambda_{111}^{XYZ} = \frac{1}{8}\log[\theta_{11(1)}/\theta_{11(2)}]。$$

因此，$\lambda_{ijk}^{XYZ} = 0$ 等价于 $\theta_{11(1)} = \theta_{11(2)}$。

8.18 分别掷两枚相同的硬币。令 X = 第一枚硬币是否正面朝上（是，否），Y = 第二枚硬币是否正面朝上，以及 Z = 掷两枚硬币的结果是否相同。利用这个例子，证明三个变量中两两边际独立并不意味着它们之间相互独立。

8.19 对于三个分类变量，X，Y 和 Z：

a. 当 Y 联合独立于 X 和 Z 时，证明在给定 Z 后 X 和 Y 条件独立。

b. 证明 X，Y 和 Z 的相互独立性意味着 X 和 Y 既边际独立也条件独立。

c. 当 X 独立于 Y 并且 Y 独立于 Z 时，能推出 X 独立于 Z 吗？加以说明。

d. 当任意一对变量都条件独立时，说明为什么不存在三维交互项。

8.20 假设在给定 Z 后 X 和 Y 条件独立，并且 X 和 Z 边际独立。

a. 证明 X 联合独立于 Y 和 Z。

b. 证明 X 和 Y 边际独立。

c. 证明如果 X 和 Z 条件独立（而不是边际独立），那么 X 和 Y 仍然边际独立。

8.21 对于所有的 i,j,k，一个 $2\times2\times2$ 表格满足 $\pi_{i++} = \pi_{+j+} = \pi_{++k} = \frac{1}{2}$。给出一个 $\{\pi_{ijk}\}$ 满足以下模型但不满足比此更简单的模型的例子：(a)(X,Y,Z)，(b)(XY,Z)，(c)(XY,YZ)，(d)(XY,XZ,YZ)，(e)(XYZ)。

8.22 假定对于一个 $2\times2\times2$ 表格，模型(XY,XZ,YZ)成立，并且在 Z 的两个取值水平上 XY 的共同条件对数发生比之比为正。如果 XZ 和 YZ 的条件对数发生比之比均为正或均为负，证明 XY 的边际发生比之比大于它们的条件发生比之比。因此，在 XY 的关联上不可能出现辛普森悖论。

8.23 证明关于 T 维表格的一般性对数线性模型包括 2^T 项（*提示：*它包括一个截距项，$\binom{T}{1}$ 个单维项，$\binom{T}{2}$ 个二维项，……）。

8.24 所有 T 个结果变量均为二分变量。对于虚拟变量 $\{z_1,\cdots,z_T\}$，满足相互独立性的对数线性模型具有以下形式：

$$\log\mu_{z_1,\cdots,z_T} = \lambda_1 z_1 + \cdots + \lambda_T z_T。$$

给出一般性对数线性模型的表达式（Cox，1972）。

8.25 考虑关于 W,X,Y,Z 的交叉列联表。

a. 说明为什么模型(WXZ,WYZ)是在 X 和 Y 满足条件独立的情况下最一般化的对数线性模型。

b. 用符号表示出 X 和 Y 条件独立，并且不存在三维交互项的模型。

8.26 对于一个关于二分结果变量 Y 的四维表格，给出在下列情况下等价的对数线性模型和 Logit 模型：

a. A,B 和 C 对 Y 的主效应。

b. A 和 B 对 Y 的效应存在交互项，C 对 Y 存在主效应。

c. 当 Y 是定类变量时，利用基线类别 Logit 模型重做(a)部分。

8.27 对于行变量为定序变量且赋值为$\{x_i\}$的 3×3 表格，指出以下模型在式 8.8 广义对数线性模型中所对应的项：(a) $\text{logit}[P(Y\leqslant j)]=\alpha_j+\beta x_i$，(b) $\log[P(Y=j)/P(Y=3)]=\alpha_j+\beta_j x_i$。

8.28 考虑关于二维表格的独立性模型，推导出它的最小充分统计量、似然方程、拟合值，以及残差自由度。

8.29 考虑关于 $I\times J$ 表格的对数线性模型，$\log\mu_{ij}=\lambda+\lambda_i^X$，证明 $\hat{\mu}_{ij}=n_{i+}/J$ 且它的残差自由度为 $df=I(J-1)$。

8.30 给出模型(XZ,YZ)的对数似然函数 L。计算$\partial L/\partial\lambda$ 并证明它意味着 $\hat{\mu}_{+++}=n$。证明$\partial L/\partial\lambda_i^X=n_{i++}-\mu_{i++}$。相似地，对似然函数求导来获得相应的似然方程组。证明式 8.23 和式 8.24 隐含着其他的方程组，所以这些方程组决定了最大似然估计值。

8.31 对于模型(XY,Z)，推导：(a)最小充分统计量，(b)似然方程，(c)拟合值，(d)拟合检验的残差自由度。

8.32 考虑标示为(XZ,YZ)的对数线性模型。

a. 对于给定的 k，证明$\{\hat{\mu}_{ijk}\}$等于在 Z 的取值为 k 时对 X 和 Y 进行独立性检验的拟合值。

b. 证明对这个模型进行拟合检验的皮尔逊和似然比统计量具有 $X^2=\sum X_k^2$ 的形式，其中 X_k^2 检验当 Z 取值为 k 时 X 和 Y 之间的独立性。

8.33 验证表 8.14 给出的关于模型(XY,Z)、(XY,YZ)以及(XY,XZ,YZ)的自由度的值。

8.34 验证对数线性模型(GLS,GI,LI,IS)等价于 Logit 模型(8.16)。证明 S 对 I 的效应的条件对数发生比之比在 Logit 模型中等于 $\beta_1^S-\beta_2^S$，在对数线性模型中等于 $\lambda_{11}^{IS}+\lambda_{22}^{IS}-\lambda_{12}^{IS}-\lambda_{21}^{IS}$。

表 8.22 习题 8.35 的数据[a]

模型	期望频数的估计	残差自由度
(W,X,Y,Z)	$n_{h+++}n_{+i++}n_{++j+}n_{+++k}/n^3$	$HIJK-H-I-J-K+3$
(WX,Y,Z)	$n_{hj++}n_{++j+}n_{+++k}/n^2$	$HIJK-HI-J-K+2$
(WX,WY,Z)	$n_{hj++}n_{h+j+}n_{+++k}/n_{h+++}n$	$HIJK-HI-HJ-K+H+1$
(WX,YZ)	$n_{hi++}n_{++jk}/n$	$(HI-1)(JK-1)$
(WX,WY,XZ)	$n_{hi++}n_{h+j+}n_{+i+k}/n_{h+++}n_{+i++}$	$HIJK-HI-HJ-IK+H+I$
(WX,WY,WZ)	$n_{hi++}n_{h+j+}n_{h++k}/(n_{h+++})^2$	$HIJK-HI-HJ-HK+2H$
(WXY,Z)	$n_{hij+}n_{+++k}/n$	$(HIJ-1)(K-1)$
(WXY,WZ)	$n_{hij+}n_{h++k}/n_{h+++}$	$H(IJ-1)(K-1)$
(WXY,WXZ)	$n_{hij+}n_{hi+k}/n_{hi++}$	$HI(J-1)(K-1)$

a W,X,Y,Z 的类别数量由 H,I,J,K 表示。有关其他模型的估计可根据对称性推出。

8.35 表 8.22 给出了在四维表格中存在直接估计的模型的拟合值。

a. 利用 Birch 的结果来验证该表中关于模型(W,X,Y,Z)的信息是正确的。验证该模型的残差自由度。

b. 利用复合变量(composite variables)以及有关二维表格的相应结果(模型(WXY, WZ)如给定 W,Z 独立于复合变量 XY),给出关于模型(WX,YZ),(WXY,Z),(WXY,WZ)和(WXY,WXZ)的相应估计和自由度公式。

8.36 某个 T 维表格$\{n_{ab\cdots t}\}$在第 i 个维度上具有 I_i 个类别。

a. 求相互独立性模型的最小充分统计量、单元格概率的最大似然估计以及残差自由度。

b. 求包括所有二维关联项但不包括任何三维交互项的分级模型的最小充分统计量和残差自由度。

8.37 考虑关于 $2\times2\times2$ 表格的对数线性模型(X,Y,Z)。

a. 将模型表示为 $\log \boldsymbol{\mu}=\mathbf{X}\boldsymbol{\beta}$ 的形式。

b. 证明似然方程 $\mathbf{X}'\mathbf{n}=\mathbf{X}'\hat{\boldsymbol{\mu}}$ 在单变量的边际分布中令$\{n_{ijk}\}$等于$\{\hat{\mu}_{ijk}\}$。

8.38 分别对模型(a)(X,YZ)和(b)(XZ,YZ)使用 IPF 算法,证明在一个循环内就可以得到最大似然估计。

8.39 给定目标的行总计$\{r_i>0\}$和列总计$\{c_j>0\}$:

a. 说明如何通过 IPF 算法去调整样本比例$\{p_{ij}\}$以获得这些总计,并同时保持样本发生比之比不变。

b. 指出如何求得满足这些总计并且所有的局部发生比之比等于 $\theta>0$ 的单元格比例(*提示*:令第一行和第一列的所有单元格的初始值都等于1.0。这在所有局部发生比之比等于 θ 的条件下决定了其他单元格的初始值)。

c. 说明单元格比例是如何由边际比例和局部发生比之比决定的。

8.40 参考第 8.6.3 节中 Birch 的结果。证明随着 $\log \mu_i\to\pm\infty$,L 的各项分别收敛于 $-\infty$。说明为什么信息矩阵的正定性意味着,在似然函数取最大值的点似然方程的解是唯一的。

9 对数线性模型和Logit模型的构建与扩展

第5-7章介绍了 logistic 回归模型,该模型对二项分布或多项分布结果变量使用 Logit 连结。在第8章中,我们介绍了关于列联表的对数线性模型,它对服从泊松分布的单元格计数使用对数连结,其中,我们还在第8.5.3节讨论了二者之间的等价性。在本章中,我们探讨在列联表的情况下,对这些模型的构建和扩展。

在第9.1节中,我们介绍利用绘图表现模型的关联和条件独立模式的方法。第9.2节讨论如何选择和比较对数线性模型。有关检查模型的诊断方法,如残差等,将在第9.3节进行介绍。

第8章所介绍的对数线性模型将所有变量都视为定类变量。第9.4节介绍关于定序变量之间关联的对数线性模型。在第9.5节、第9.6节中,我们给出利用参数来替代固定赋值的一般化模型。在本章的最后一节,我们讨论在列联表存在稀疏数据的情况下可能出现的问题。

9.1 关联图与可合并性

可以通过绘图的方法来描述对数线性模型中的关联,这种图示法能够直观展现满足条件独立的各对变量,因而有助于我们理解模型的含义。这里的内容部分借鉴了 Darroch 等(1980)的成果,他们利用了数图理论来描述具有条件独立性结构的对数线性模型(称为*图示模型*(*graphical models*))。

9.1.1 关联图

关联图(*association graph*)由一组顶点组成,每个顶点代表一个变量。两个变量之间的连接线表示这两个变量之间存在条件关联。例如,对数线性模型(WX,WY,WZ,YZ)缺少 XY 和 XZ 项。它假定,以其他两个变量为条件,X 和 Y 以及 X 和 Z 之间满足独立性。图9.1给出了这个模型的关联图:四个变量形成了图的顶点,四条线分别代表了成对变量之间的条件关联。在 X 和 Y 以及 X 和 Z 之间不存在连接线,表明这两对变量条件独立。

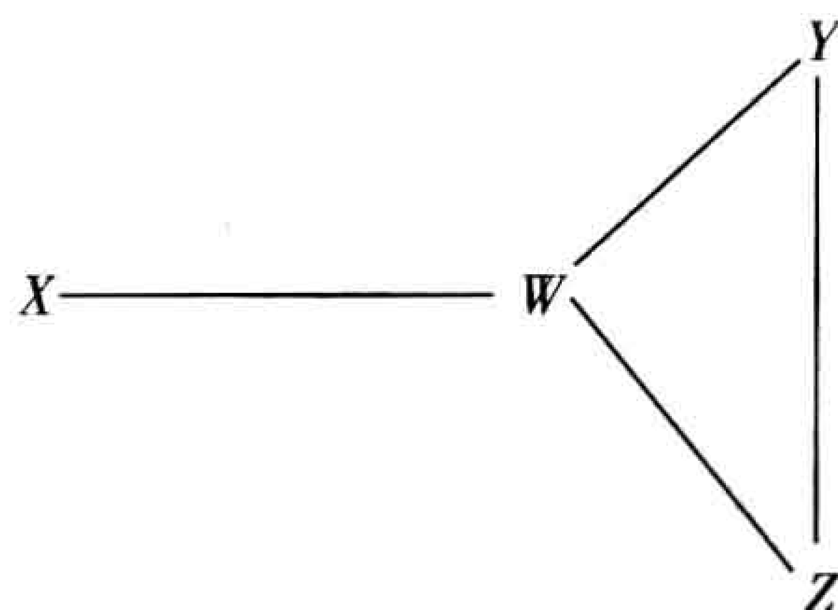


图9.1 模型(WX,WY,WZ,YZ)的关联图

具有相同的成对关联模式的两个对数线性模型对应的关联图也一样。例如,上面的关联图对模型(WX,WYZ)也适用,与前面的模型相比,它增加了一个三维交互项 WYZ。

在关联图中,*路径*(*path*)是指从一个变量到另一个变量之间的一系列连接线。如果

所有连接 X 和 Y 的路径都必须通过另外一个变量集，那么我们就说两个变量 X 和 Y 是*分离的*（*separated*）。例如，在图 9.1 中，W 分离了 X 和 Y，因为所有连接 X 和 Y 的路径都经过 W。变量集 $\{W,Z\}$ 也分离了 X 和 Y。一个基本结论是，在分离两个变量的变量集中，给定其中的*任意*子集，这两个变量之间条件独立（Kreiner，1987；Whittaker，1990：67）。因此，不仅在给定 W 和 Z 后 X 和 Y 条件独立，而且在只给定 W 时 X 和 Y 也条件独立。相似地，在只给定 W 时，X 和 Z 条件独立。

9.1.2 三维列联表的可合并性

在第 2.3.3 节中，我们指出分表中的条件关联往往不同于边际关联。但是，在一定的*可合并性条件*（*collapsibility conditions*）下，它们是相同的。

对于三维表格，如果 Z 和 X 条件独立或者 Z 和 Y 条件独立，那么 XY 的边际发生比之比与条件发生比之比完全相同。

该条件是说，作为控制变量的（Z）要与 X 或 Y 或两者都条件独立。对数线性模型（XY，YZ）和（XY，XZ）都满足这些条件。因此，对于具有关联图为

$$X\text{——}Y\text{——}Z \text{ 和 } Y\text{——}X\text{——}Z$$

的模型，甚至更简单的模型，在分表中所拟合的 XY 发生比之比与在边际表中相同。但是对于具有关联图为

$$X\text{——}Z\text{——}Y$$

的模型，其中一条线将 Z 与 X 和 Y 连在一起，上述结论不成立。可以通过关于模型（XY，YZ）和（XY，XZ）的有关公式直接对此进行证明（习题 9.26）。

我们以第 8.2.4 节中有关学生调查的数据（表 8.3）为例来加以说明，其中 A = 饮酒，C = 抽烟，M = 吸食大麻。模型（AM，CM）设定，在给定 M 后，AC 满足条件独立。它所对应的关联图为

$$A\text{——}M\text{——}C。$$

考虑 AM 的关联。由于 C 条件独立于 A，所拟合的 AM 条件发生比之比与对 C 进行合并后所拟合的 AM 边际发生比之比相等。根据表 8.5 的数据，它们都等于 61.9。相似地，CM 的关联也具有可合并性。AC 关联不可以合并，因为在模型（AM，CM）中，M 与 A 和 C 都条件相依。因此，尽管 A 和 C 条件独立，它们之间可能在边际上不独立。事实上，由表 8.5 可知，这个模型所拟合的 AC 边际发生比之比等于 2.7。

对于模型（AC，AM，CM），没有一对变量之间条件独立，因而都不满足可合并性条件。表 8.5 表明，对于这个模型，每对变量之间所拟合的边际关联和条件关联都相差很大。当一个模型包括所有的二维效应时，对任何一个变量进行合并都可能导致效应的变化。

9.1.3 可合并性与 Logit 模型

这些可合并性条件同样适用于 Logit 模型。例如，假定在 K 个中心（Z）进行的一项临床试验，研究一个二分干预变量 X（$x_1=1$，$x_2=0$）与二分结果变量 Y 之间的关联。相应的 Logit 模型为

$$\text{Logit}[P(Y=1) \mid X=i,Z=k] = \alpha + \beta x_i + \beta_k^Z,$$

对每个中心来说，干预效应 β 都相同。由于这个模型对应于对数线性模型（XY，XZ，YZ），在将 $2\times2\times K$ 表格对 K 个中心进行合并后，这个效应可能会发生变化。同样，所估计的 XY 条件发生比之比 $\exp(\hat{\beta})$，一般不等于 2×2 边际表中的样本发生比之比。

接下来，考虑不包括中心效应的较简单模型，

$$\text{Logit}[P(Y=1 \mid X=i, Z=k)] = \alpha + \beta x_i。$$

给定一种干预方式,结果为成功的概率对每个中心都一样。这个模型满足可合并性条件,因为它表明,在给定 X 后,Z 条件独立于 Y。这个模型等价于对数线性模型(XY,XZ),其中 XY 的关联是可合并的。因而,当中心效应可以忽略并且较简单的模型能很好地拟合数据时,所估计的干预效应近似等于 XY 的边际发生比之比。

9.1.4　多维表格的可合并性与关联图

Bishop 等(1975,p.47)给出了关于多维表格的参数可合并性条件:

> 在一个关于多维表格的模型中,假定将变量分割为三个互不重合的子集——A,B,C,以使得 B 分离了 A 和 C。在按照 C 中的变量将表格进行合并后,关于 A 中变量的参数以及 A 和 B 中变量之间的关联参数不变。

我们利用模型(WX,WY,WZ,YZ)(图 9.1)加以说明。令 $A=\{X\}$,$B=\{W\}$,$C=\{Y,Z\}$。由于模型不包括 XY 和 XZ 项,所有连接子集 A 和子集 C 的参数都等于零,并且 B 分离了 A 和 C。如果我们对 Y 和 Z 进行合并,WX 的关联不会发生变化。类似地,令 $A=\{Y,Z\}$,$B=\{W\}$,$C=\{X\}$。那么,在对 X 进行合并后,关于 W,Y 以及 Z 之间的条件关联保持不变。

这一结果表明,当任一变量独立于其他所有变量时,对其进行合并不会影响模型中的其他项。例如,在模型(WX,WY,XY,Z)与模型(WX,WY,XY)中,关于 W,X 和 Y 之间的关联是一样的。

当子集 B 包含多个变量时,尽管在对子集 C 进行合并后参数的值不变,对这些参数的最大似然估计可能会略有区别。更强的可合并性条件要求这些估计也必须相等。如果模型中包含了子集 B 中各变量之间的最高阶项,强可合并性成立。Asmussen 和 Edwards(1983)讨论了这一特性,它与表格的*可分解性*(*decomposability*)有关(注解 8.2)。

9.2　模型选择与比较

选择和比较对数线性模型的策略与第 6.1 节所介绍的关于 logistic 回归的情况类似。一个模型既应当足够复杂以保证能够很好地拟合数据,又应当相对简单以便于解释。值得强调的是,模型是对数据的修匀,而不是过度拟合。

9.2.1　模型选择的考虑

具有潜在价值的模型通常是所有可能模型的一个很小的子集。一项为了通过验证性分析回答某些问题所设计的研究,可能只需要对比包括或不包括某些特定项的模型的拟合结果。同时,模型应当体现出结果变量和解释变量之间的区别。构建模型的过程应当突出强调与结果变量有关的项以及将解释变量与结果变量联系起来的项。模型应当包括解释变量之间的最一般的交互项。从似然方程来看,这具有在解释变量的取值组合点上使所有的拟合总计等于样本总计的效果。这是一种自然的结果,因为我们通常都将这些样本总计视为固定的。与此相联系,某些边际总计在抽样设计中往往是固定的。任何潜在的模型都应当把这些边际总计作为充分统计量包括进来,从而在似然方程中保证所拟合的总计与它们相等。

以表 8.8 为例,我们将 I = 车祸受伤和 S = 使用安全带作为结果变量,将 G = 性别

和 L = 地点作为解释变量。这时,我们将在每个 G 和 L 的取值组合点上的 $\{n_{g+l+}\}$ 视为固定的。例如,20 629 名女性在城区发生车祸,因而拟合的计数也应当在城区有 20 629 名女性。为了保证这一结果,对数线性模型应该包含 GL 项,在似然方程中这意味着 $\{\hat{\mu}_{g+l+}=n_{g+l+}\}$。因此,模型应当至少与($GL,S,I$)一样复杂,并同时关注 G 和 L 对 S 和 I 的效应以及 SI 之间的关联。

如果 S 也是一个解释变量而只有 I 是结果变量,那么 $\{n_{g+ls}\}$ 应当是固定的。当只有一个分类结果变量时,相应的对数线性模型对应于一个关于该结果变量的 Logit 模型。这时,应当考虑使用 Logit 模型而不是对数线性模型,因为主要的关注点是描述解释变量对结果变量的效应。

探索性研究在潜在的模型中寻找适当模型,这个过程会提供很多关于变量间关联和交互模式的线索。一种方法是首先拟合只包括单维项的模型,接着在模型中加入二维项,然后继续加入三维项,以此类推。这样进行模型拟合常常能够找到很小的一组拟合结果很好的模型。在第 8.4.2 节中,我们对车祸数据的分析就采取了这种策略。一些在可能的模型中自动寻找的方法,比如后向剔除法,可能也会有所帮助,但是在使用这些方法时一定要慎重,因为通过这类方法所得出的模型并不一定有意义。

9.2.2 关于代顿学生调查数据的模型构建

在第 8.2.4 和第 8.3.2 节中,我们分析了一个高三学生样本有关饮酒(A)、抽烟(C),以及吸食大麻(M)的数据。该研究所收集的信息还包括学生的性别(G)和种族(R)。表 9.1 给出了相应的五维表格。在选择模型的过程中,我们将 A,C 和 M 作为结果变量,将 G 和 R 作为解释变量。因此,模型中应当包括 GR 项,以保证所拟合的 GR 边际总计等于样本的边际总计。

表 9.1 高三学生饮酒、抽烟和吸食大麻的情况

饮酒	抽烟	吸食大麻							
		种族 = 白人				种族 = 其他			
		女生		男生		女生		男生	
		是	否	是	否	是	否	是	否
是	是	405	268	453	228	23	23	30	19
	否	13	218	28	201	2	19	1	18
否	是	1	17	1	17	0	1	1	8
	否	1	117	1	133	0	12	0	17

来源:Harry Khamis,Wright State University。

表 9.2 给出了几个模型的拟合优度检验结果。由于很多单元格计数偏小,G^2 对卡方的近似可能较差,但是这个指标仍然可以用来对不同模型进行比较。表中所列出的第一个模型只包含了 GR 关联,并假定所有其他九对关联都条件独立。毫不奇怪,它拟合得非常差。相比较而言,模型 2 包括了所有的二维项,看起来拟合得不错。模型 3 包括了所有的三维交互项,也拟合得很好,但与模型 2 相比,拟合结果的改进有限(二者 G^2 之差等于 $15.3-5.3=10.0$,基于 $df=16-6=10$ 个自由度)。因此,我们考虑不包括三维交互项的模型。以模型 2 作为起点,考虑删除一些二维项。我们使用后向剔除法,在重新拟合模型时,依次剔除那些相对应的 G^2 增加最小的项。

表 9.2 展示了模型选择的过程。9 个成对关联项(GR 项除外)是可能从模型 2 中剔除的候选项,在表中由模型 4a 至模型 4i 表示。与模型 2 相比较,删除 CR 项(即模型 4g)

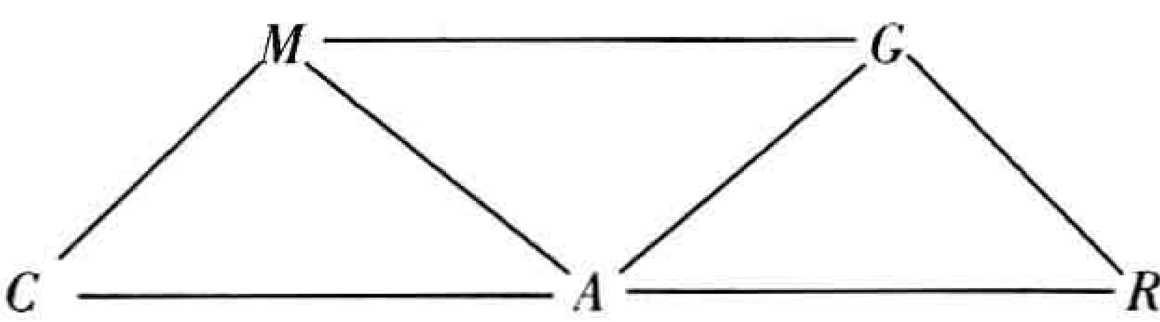
时 G^2 增加最小。它的增加值为 15.8 - 15.3 = 0.5，自由度为 df = 17 - 16 = 1，所以去除该项是可以接受的。在剔除该项后，下一个 G^2 的最小增加值发生在删除 *CG* 项（模型 5）时，这时模型对应的 G^2 = 16.7，df = 18，也即在自由度变动 df = 1 的情况下 G^2 变动了 0.9。接着，剔除 *MR* 项（模型 6）后 G^2 = 19.9，df = 19，即基于 df = 1，G^2 变动了 3.2。

表 9.2　关于表 9.1 数据的对数线性模型的拟合优度检验

模　型[a]	G^2	df
1. 相互独立性 + *GR*	1 325.1	25
2. 同性质关联	15.3	16
3. 所有的三维交互项	5.3	6
4a. (2) - *AC*	201.2	17
4b. (2) - *AM*	107.0	17
4c. (2) - *CM*	513.5	17
4d. (2) - *AG*	18.7	17
4e. (2) - *AR*	20.3	17
4f. (2) - *CG*	16.3	17
4g. (2) - *CR*	15.8	17
4h. (2) - *GM*	25.2	17
4i. (2) - *MR*	18.9	17
5. (*AC*, *AM*, *CM*, *AG*, *AR*, *GM*, *GR*, *MR*)	16.7	18
6. (*AC*, *AM*, *CM*, *AG*, *AR*, *GM*, *GR*)	19.9	19
7. (*AC*, *AM*, *CM*, *AG*, *AR*, *GR*)	28.8	20

a *G* 性别；*R* 种族；*A*，饮酒；*C*，抽烟，*M*，吸食大麻。

进一步剔除模型中的剩余项都会对拟合结果产生严重影响。例如，去除 *AG* 项会使 G^2 增加 5.3，在 df = 1 下对应的 *P* 值为 0.02。尽管我们不应该因为数据显示如此，就过分看重这些 *P* 值，但是看起来不再做进一步的剔除是比较安全的（有关在多重检验中调整 *P* 值的方法，参见：Westfall and Wolfinger(1997)、Westfall and Young(1993)）。模型 6，由（*AC*, *AM*, *CM*, *AG*, *GM*, *GR*）来表示，它对应的关联图为：



连接 *C* 和 {*G*, *R*} 的每条路径都要经过 {*A*, *M*} 中的变量。模型表明，给定饮酒和吸食大麻的结果，抽烟行为独立于性别和种族。按照解释变量性别和种族对数据进行合并，拟合关于 *C* 和 *A* 以及 *C* 和 *M* 之间的条件关联，结果与在第 8.2.4 节所拟合的模型（*AC*, *AM*, *CM*）相同。

在模型 6 中剔除 *GM* 项就得到了表 9.2 中的模型 7。它的关联图显示，*A* 分离了 {*G*, *R*} 和 {*C*, *M*}。因此，在对 *G* 和 *R* 进行合并后，模型 7 中关于 *A*, *C* 以及 *M* 之间的所有成对条件关联都与模型（*AC*, *AM*, *CM*）相一致。事实上，考虑到巨大的样本规模，模型 7 的拟合结果并不差（G^2 = 28.8，df = 20），它对应的样本差异指数 $\hat{\Delta}$ = 0.036。所以，在分析所关注的主要变量之间的关联时，可以考虑对性别和种族进行合并。包括所有 5 个变量的模型的优势在于，它估计了性别和种族对这些结果变量的效应，尤其是种族和性别对饮酒的效应以及性别对吸食大麻的效应。

9.2.3 比较对数线性模型的统计量

考虑两个对数线性模型 M_1 和 M_0，其中 M_0 是 M_1 的一个特例。根据第 4.5.4 和第 5.4.3节的结果，以 M_1 作为备择模型对 M_0 进行检验的似然比统计量为 $G^2(M_0|M_1) = G^2(M_0) - G^2(M_1)$。我们可以利用这个统计量来比较两个模型。

记 $\mathbf{n}$ 为由观察到的单元格计数 $\{n_i\}$ 所形成的列向量，令 $\hat{\boldsymbol{\mu}}_0$ 和 $\hat{\boldsymbol{\mu}}_1$ 分别为模型 M_0 和 M_1 的拟合值 $\{\hat{\mu}_{0i}\}$ 和 $\{\hat{\mu}_{1i}\}$ 的向量。较简单模型所对应的偏离度 $G^2(M_0)$ 可以分割为

$$G^2(M_0) = G^2(M_1) + G^2(M_0 \mid M_1)。 \tag{9.1}$$

如同 $G^2(M)$ 测量模型 M 的拟合值与 $\mathbf{n}$ 之间的距离一样，$G^2(M_0|M_1)$ 测量拟合值 $\hat{\boldsymbol{\mu}}_0$ 与拟合值 $\hat{\boldsymbol{\mu}}_1$ 之间的距离。在这个意义上，因式分解公式 9.1 具有一定的正交性：$\mathbf{n}$ 与 $\hat{\boldsymbol{\mu}}_0$ 之间的距离等于 $\mathbf{n}$ 与 $\hat{\boldsymbol{\mu}}_1$ 之间的距离再加上 $\hat{\boldsymbol{\mu}}_1$ 与 $\hat{\boldsymbol{\mu}}_0$ 之间的距离。

用以比较模型的统计量等于

$$\begin{aligned} G^2(M_0 \mid M_1) &= 2\sum_i n_i \log(n_i/\hat{\mu}_{0i}) - 2\sum_i n_i \log(n_i/\hat{\mu}_{1i}) \\ &= 2\sum_i n_i \log(\hat{\mu}_{1i}/\hat{\mu}_{0i})。 \end{aligned} \tag{9.2}$$

两个对数线性模型具有公式 8.17 的矩阵形式，即

$$\log \boldsymbol{\mu}_0 = \mathbf{X}_0 \boldsymbol{\beta}_0 \quad 和 \quad \log \boldsymbol{\mu}_1 = \mathbf{X}_1 \boldsymbol{\beta}_1。$$

由于模型 M_0 比模型 M_1 简单，它可以表述为 $\log \boldsymbol{\mu}_0 = \mathbf{X}_0\boldsymbol{\beta}_0 = \mathbf{X}_1\boldsymbol{\beta}_1^*$，其中，$\boldsymbol{\beta}_1^*$ 与 $\boldsymbol{\beta}_0$ 中相对应的元素相等，对于包括在 $\boldsymbol{\beta}_1$ 而不包括在 $\boldsymbol{\beta}_0$ 中的参数，$\boldsymbol{\beta}_1^*$ 中相应加入 0 元素。这样，由式 9.2 可得，

$$\begin{aligned} G^2(M_0 \mid M_1) &= 2\mathbf{n}'(\log \hat{\boldsymbol{\mu}}_1 - \log \hat{\boldsymbol{\mu}}_0) = 2\mathbf{n}'[\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^*] \\ &= 2\hat{\boldsymbol{\mu}}_1[\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^*] = 2\hat{\boldsymbol{\mu}}_1'(\log \hat{\boldsymbol{\mu}}_1 - \log \hat{\boldsymbol{\mu}}_0) \\ &= 2\sum \hat{\mu}_{1i} \log\left(\frac{\hat{\mu}_{1i}}{\hat{\mu}_{0i}}\right) \end{aligned} \tag{9.3}$$

其中，根据模型 M_1 的似然方程 $\mathbf{n}'\mathbf{X}_1 = \hat{\boldsymbol{\mu}}'_1\mathbf{X}_1$（回顾式 8.22），$\mathbf{n}$ 由 $\hat{\boldsymbol{\mu}}_1$ 来替代。式 9.3 统计量具有与 $G^2(M_0)$ 相同的形式，但是它利用 $\{\hat{\mu}_{1i}\}$ 取代了所观察到的数据的位置。注意，$G^2(M_0)$ 是当 M_1 为饱和模型时 $G^2(M_0|M_1)$ 的一个特例。

两个皮尔逊统计量之差 $X^2(M_0) - X^2(M_1)$ 不再具有皮尔逊统计量的形式，它的值甚至不一定是非负的。更适于进行模型比较的皮尔逊统计量为

$$X^2(M_0 \mid M_1) = \sum \frac{(\hat{\mu}_{1i} - \hat{\mu}_{0i})^2}{\hat{\mu}_{0i}}。 \tag{9.4}$$

它具有常用的皮尔逊统计量的形式，只是用 $\{\hat{\mu}_{1i}\}$ 取代了 $\{n_i\}$。式 9.3 和式 9.4 统计量只通过模型的拟合值才跟数据有关，因而，它们也只取决于模型 M_1 的充分统计量。

当模型 M_0 成立时，$G^2(M_0)$ 和 $G^2(M_1)$ 渐近服从卡方分布，进而 $G^2(M_0|M_1)$ 也渐近服从自由度为两模型自由度之差的卡方分布。Haberman（1977a）表明，$G^2(M_0|M_1)$ 和 $X^2(M_0|M_1)$ 具有相同的零分布大样本特性，即便是在稀疏数据的情况下也是如此（在一定条件下，随着 n 的增加，二者之差依概率收敛于 0）。当模型 M_1 成立但 M_0 不成立时，$G^2(M_1)$ 仍然渐近服从卡方分布，但是另外两个统计量会随着 n 的增加而无限增大。

9.2.4 模型比较的卡方分割

方程式 9.1 利用了有关 df > 1 的卡方统计量可以进行分割的特性。在线性 Logit 或

线性概率模型中对定序预测变量进行趋势检验时(第 5.3.5 节)以及在定序结果变量的累积 Logit 模型中(第 7.2 节),我们曾经应用过这样的分割。更一般地,这一特性适用于通过一组嵌套模型来检验一系列假设。用来比较每对模型的单独检验之间渐近独立。

例如,在第 3.3.3 节中通过 $J-1$ 个模型将 $2\times J$ 表格进行的卡方分解,表明了分割 G^2 的合理性。对于 $j=2,\cdots,J$,令 M_j 表示满足以下条件的模型,

$$\theta_i = (\mu_{1i}\mu_{2,i+1})/(\mu_{1,i+1}\mu_{2i}) = 1, \quad i = 1,\cdots,j-1。$$

就 M_j 而言,包含第 1 列到第 j 列的 $2\times j$ 表格都满足独立性。在完整的 $2\times J$ 表格中,模型 M_J 表示独立性模型。对于任意的 $h>j$,模型 M_h 是模型 M_j 的一个特例。由式 9.2 可得,

$$\begin{aligned} G^2(M_J) &= G^2(M_J \mid M_{J-1}) + G^2(M_{J-1}) \\ &= G^2(M_J \mid M_{J-1}) + G^2(M_{J-1} \mid M_{J-2}) + G^2(M_{J-2}) \\ &= \cdots = G^2(M_J \mid M_{J-1}) + \cdots + G^2(M_3 \mid M_2) + G^2(M_2)。\end{aligned}$$

根据式 9.3,$G^2(M_j|M_{j-1})$ 具有 G^2 形式,其中以模型 M_{j-1} 的拟合值取代了所观察到的数据。将两个模型的拟合值都代入式 9.3 可得,$G^2(M_j|M_{j-1})$ 与检验 2×2 表格独立性的 G^2 相同;该表格中,第一列等于合并了原来表格的第 1 至第 $j-1$ 列,而第二列则是原来表格的第 j 列。

当事先计划对多个模型进行比较时,同步检验程序(simultaneous test procedures)可以减少将单纯反映随机变动的样本效应当作重要发现的概率。这些程序使用的是调整后的显著性水平。对于一组嵌套模型之间的 s 个检验,当每个检验的显著性水平为 $1-(1-\alpha)^{1/s}$ 时,总体检验的渐近 P 值才满足 P(第一类错误)$\leqslant\alpha$(Goodman,1969a)。例如,假定我们检验模型(WXZ,WY,XY,ZY)的拟合,将其与模型(WX,WZ,XZ,WY,XY,ZY)进行比较,接着再与模型(WX,WZ,XZ,WY,ZY)进行比较。对于这组 $s=3$ 的检验,为了保证总体的 $\alpha=0.05$,每个检验所使用的显著性水平应为 $1-(0.95)^{1/3}=0.017$。

9.2.5 等价的边际独立性和条件独立性检验

当两个模型都存在直接估计时,利用 $G^2(M_0|M_1)$ 进行的检验可以大大简化。这种情况下,具有独立性关系的模型必然保证数据的可合并性。条件独立性检验与对边际表格进行独立性检验的结果完全相同。Sundberg(1975)证明,当两个存在直接估计的模型 M_0 和 M_1 除了一个成对关联项外完全相同时,$G^2(M_0|M_1)$ 与对该关联相应的两个变量所形成的边际表格进行独立性检验的 G^2 相等。Bishop(1971)和 Goodman(1970,1971b)对此进行了讨论。

例如,$G^2[(X,Y,Z)|(XY,Z)]$ 检验在模型(XY,Z)中 $\lambda^{XY}=0$。因此,它检验在 X 和 Y 联合独立于 Z 的假定下,XY 的条件独立性。利用这两组拟合值,由式 9.3 可得,该统计量等于

$$\begin{aligned} &2\sum_i\sum_j\sum_k \frac{n_{ij+}n_{++k}}{n}\log\frac{n_{ij+}n_{++k}/n}{n_{i++}n_{+j+}n_{++k}/n^2} \\ &= 2\sum_i\sum_j n_{ij} + \log\frac{n_{ij+}}{n_{i++}n_{+j+}/n}, \end{aligned}$$

其等于在 XY 的边际表格中进行独立性检验的 $G^2[(X,Y)]$。这个结果并不奇怪,因为对于模型(XY,Z),可合并性条件意味着 XY 的边际关联与其条件关联相同。

9.3 模型检查与诊断

利用 $G^2(M_0 \mid M_1)$ 进行模型比较有助于发现，一个额外的项是否可以改善模型的拟合情况。单元格残差则提供了检查模型对某个特定单元格拟合不足的一种工具。

9.3.1 对数线性模型的残差

在第 4.5.5 节我们指出，独立性模型的残差（第 3.3.1 节）可以扩展到任意泊松分布广义线性模型。对于列联表中观察计数为 n_i、拟合计数为 $\hat{\mu}_i$ 的单元格 i，*皮尔逊残差*（*Pearson residual*）可表示为

$$e_i = \frac{n_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}。 \tag{9.5}$$

它与皮尔逊统计量的关系为 $\sum e_i^2 = X^2$。

与二项分布模型的皮尔逊残差式 6.1 相同，$\{e_i\}$ 的渐近方差小于 1.0。它们平均约等于（残差自由度）/（单元格数量）。Haberman（1973a）将标准化皮尔逊残差定义为

$$r_i = e_i / \sqrt{1 - \hat{h}_i},$$

其中杠杆力 $\hat{h}_i$ 是所估计的帽子矩阵的对角线元素（第 4.5.5 节）。标准化皮尔逊残差渐近服从标准正态分布，因而它比皮尔逊残差更好。对于存在直接估计的对数线性模型，标准化皮尔逊残差具有封闭形式的表达式（Haberman，1978：275）。除皮尔逊残差外，也可以根据偏离度的相应信息定义其他残差（第 4.5.5 节）。

9.3.2 例子：再析学生调查的数据

在第 9.2.2 节的讨论中，我们指出，对于按表 9.1 中性别和种族对饮酒、抽烟以及吸食大麻进行交叉划分的数据，包括所有二维关联项的模型可能成立。对此模型，唯一一个取值偏大的标准化皮尔逊残差等于 3.2，该残差对应于一个观察计数为 8 而拟合值仅为 3.1 的单元格。进一步的比较表明，较简单的模型（*AC*，*AM*，*CM*，*AG*，*AR*，*GM*，*GR*）是充分的。它所存在的唯一一个取值偏大的标准化残差等于 3.3，源于它对上述单元格的拟合值仅为 2.9。不饮酒、不吸食大麻但却抽烟的非白人男性的数目在一定程度上比两个模型所预测的都大一些。考虑到巨大的样本规模以及所包括的单元格数量等因素，标准化皮尔逊残差表明两个模型的拟合都没有问题。

9.3.3 对数线性模型残差与 Logit 残差的对应关系

在第 8.5 节中，我们指出关于列联表的 Logit 模型等价于某个对数线性模型。然而，Logit 模型的皮尔逊残差不同于相应对数线性模型的皮尔逊残差。由于模型的拟合值相同，比较服从二项分布或泊松分布的第 i 个观察计数与拟合计数对应的皮尔逊残差的分子相同，但是，在分母上，Logit 模型所使用的是二项分布标准差的拟合值（参见式 6.1），而对数线性模型使用的是泊松分布标准差的拟合值（参见式 9.5）。因此，Logit 模型的皮尔逊残差大于由式 9.5 所表示的对数线性模型的相应皮尔逊残差。

通过对残差除以所估计的标准误进行标准化后，两种模型的标准化皮尔逊残差是相等的。这也是偏好标准化残差而不是普通的皮尔逊残差的另一原因。

9.4 对定序关联的模型分析

到目前为止所介绍的对数线性模型存在一个严重的缺陷——它们把所有的变量测度都作为定类尺度。无论变量类别之间的排序如何变化，模型的拟合结果都一样。对于定序尺度的变量，这些模型忽略了重要的数据信息。

以表 9.3 为例。被调查者被问及他们对男女在婚前发生性行为的态度（总是错的，几乎总是错的，有时是错的，根本没错）。同时，他们还回答了是否应当向年龄在 14～16 岁的青少年提供避孕服务（强烈不同意，不同意，同意，强烈同意）。由 I 来表示相应的独立性对数线性模型，$G^2(I)=127.6$，$df=9$。这个模型拟合得很差，但是加入普通的交互项后得到的是饱和模型，因而毫无帮助。

表 9.3　关于婚前性行为与向青少年提供避孕服务的态度

婚前性行为	向青少年提供避孕服务[a]			
	强烈不同意	不同意	同意	强烈同意
总是错的	81	68	60	38
	(42.4)[1]	(51.2)	(86.4)	(67.0)
	7.6[2]	3.1	-4.1	-4.8
	(80.9)[3]	(67.6)	(69.4)	(29.1)
几乎总是错的	24	26	29	14
	(16.0)	(19.3)	(32.5)	(25.2)
	2.3	1.8	-0.8	-2.8
	(20.8)	(23.1)	(31.5)	(17.6)
有时是错的	18	41	74	42
	(30.1)	(36.3)	(61.2)	(47.4)
	-2.7	1.0	2.2	-1.0
	(24.4)	(36.1)	(65.7)	(48.8)
根本没错	36	57	161	157
	(70.6)	(85.2)	(143.8)	(111.4)
	-6.1	-4.6	2.4	6.8
	(33.0)	(65.1)	(157.4)	(155.5)

a 1 独立性模型的拟合值；2 独立性模型的标准化皮尔逊残差；3 双线性关联模型的拟合值。

来源：1991 年综合社会调查（General Social Survery），美国民意研究中心（National Opinion Research Center）。

表 9.3 还给出了独立性模型的拟合值和标准化残差。在表格的每个角部对应的残差都很大。当两个结果变量的取值都是非常负面或非常正面时，相应单元格的样本计数比独立性模型所预测的大很多。相反，当一个非常正面而另一个非常负面时，样本计数远小于模型的拟合值。定序变量之间的列联表常常出现位于角部的单元格严重偏离独立性的情况。表 9.3 所显示的残差模式表明，模型拟合不足是因为缺乏一个反映正向趋势的项。更支持向青少年提供避孕服务的对象往往也对婚前性行为更宽容。

关于定序变量的模型通过关联项来表示这种趋势。这样的模型比独立性模型更复杂，但是还没达到饱和模型。在定类变量的饱和模型中也存在关联项和交互项，然而在定序模型中，有关的趋势检验具有更高的统计效能。

9.4.1 二维表中的双线性关联

在二维表格中，当两个变量都是定序变量时，存在一个简单模型分别对行和列使用

以下排序的赋值，行赋值为 $u_1 \leqslant u_2 \leqslant \cdots \leqslant u_I$，列赋值为 $v_1 \leqslant v_2 \leqslant \cdots \leqslant v_J$。该模型可表示为

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j, \tag{9.6}$$

其约束条件可以设为如 $\lambda_I^X = \lambda_J^Y = 0$。这个模型是式 8.2 饱和模型中令 $\lambda_{ij}^{XY} = \beta u_i v_j$ 的一个特例。它只需要一个参数来描述关联，而饱和模型却需要 $(I-1)(J-1)$ 个。

当 $\beta = 0$ 时，上述模型便简化为独立性模型。$\beta u_i v_j$ 项反映了 $\log \mu_{ij}$ 对独立性的偏离。在给定 X 的取值后，这个偏离对 Y 的赋值是线性的；在给定 Y 的取值后，它对 X 的赋值也是线性的。例如，在第 j 列，该偏离程度是 X 的一个线性函数，它具有（斜率）×（X 的赋值）的形式，其中斜率为 βv_j。由于这一特性，式 9.6 被称为*双线性关联模型*（*linear-by-linear association model*，简称为 $L \times L$）。这个模型允许在表格的角部对独立性的偏离最大。Birch（1965）、Goodman（1979a）以及 Haberman（1974b）介绍了与该模型有关的一些特殊情况。

关联的方向和强度取决于 β。当 $\beta > 0$ 时，Y 随着 X 的上升而上升。在 X 和 Y 的取值都很大或很小的单元格中，该模型拟合的期望频数比在独立性模型下的期望值大。当 $\beta < 0$ 时，Y 随着 X 的上升而下降。如果数据展示出某种正向或负向趋势，$L \times L$ 模型的拟合结果通常会比独立性模型好得多。

将行 a 和 c 与列 b 和 d 相交叉所形成的 2×2 表格直接代入式 9.6，可见模型具有

$$\log \frac{\mu_{ab}\mu_{cd}}{\mu_{ad}\mu_{cb}} = \beta(u_c - u_a)(v_d - v_b)。\tag{9.7}$$

这个对数发生比之比随着 $|\beta|$ 的增加或者两对类别间距的扩大而增大。当 $u_2 - u_1 = \cdots = u_I - u_{I-1}$ 且 $v_2 - v_1 = \cdots = v_J - v_{J-1}$ 时，模型存在非常简单的解释。例如，对于 $\{u_i = i\}$ 和 $\{v_j = j\}$，相邻行和相邻列之间的*局部发生比之比*（*local odds ratio*）（式 2.10）具有共同的值 e^β。Goodman（1979a）称之为*一致性关联*（*uniform association*）。图 9.2 画出了具有相同值的局部发生比之比。

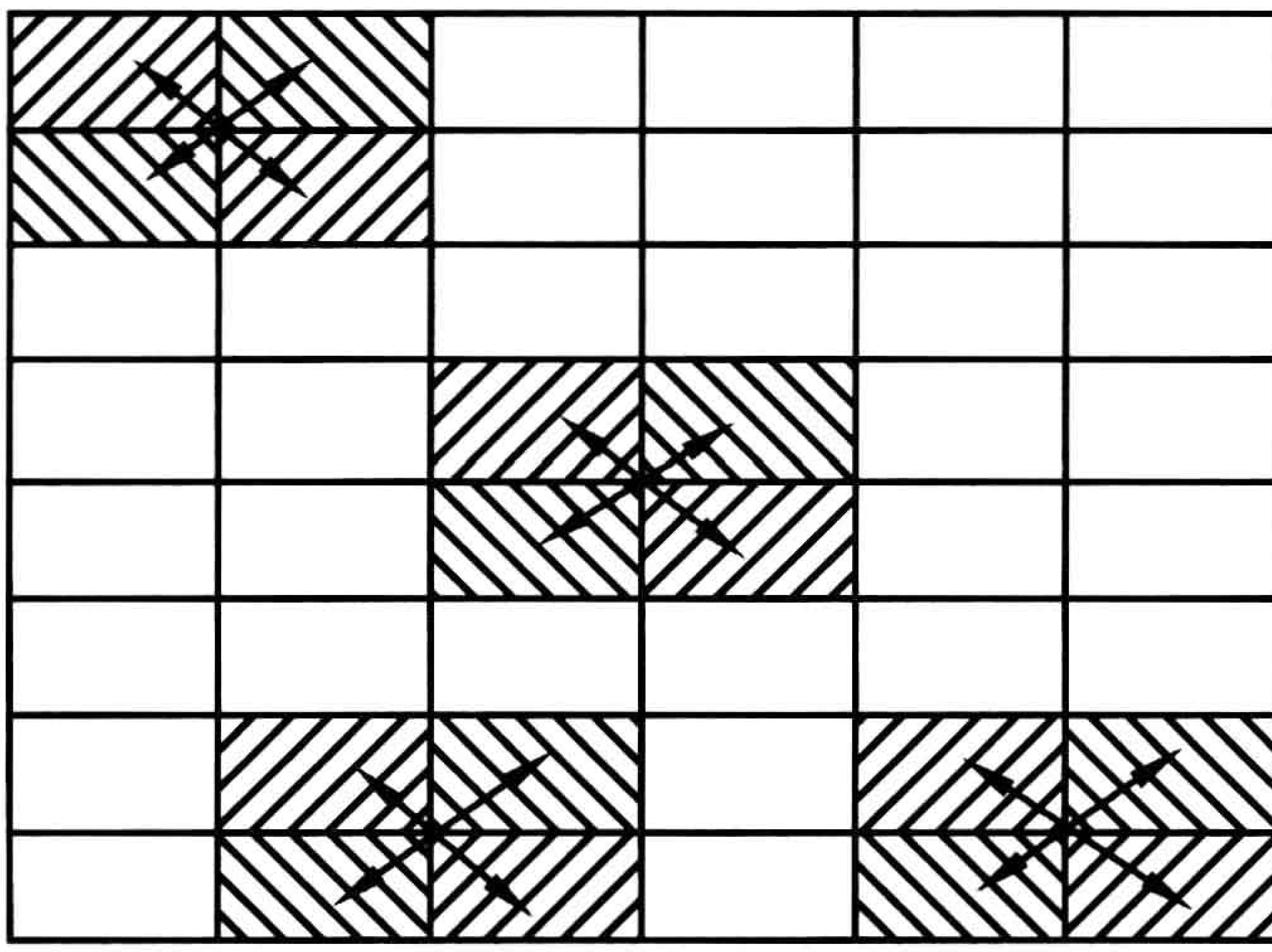


图 9.2　一致性关联模型隐含着发生比之比为常数。（注：β = 相邻行和相邻列的常数对数发生比之比）

对赋值的选取会影响关于 β 的解释。通常情况下，结果变量的刻度可视为对内在的连续尺度所进行的离散处理。这时，一种合理的赋值办法是，选取所隐含连续尺度在每个类别所有取值的中点之间的近似距离作为赋值，就像我们在第 3.4.5 节讨论的线性 Logit 模型中对饮酒量的赋值一样。有时，对这些赋值进行标准化——即减去其均值并除

以其标准差,会很有帮助。这样,

$$\sum u_i \pi_{i+} = \sum v_j \pi_{+j} = 0$$

$$\sum u_i^2 \pi_{i+} = \sum v_j^2 \pi_{+j} = 1。$$

相应地,β 表示以标准差为单位的 X 和 Y 的距离之间的对数发生比之比。当隐含的连续分布近似于二元正态分布时,$L\times L$ 模型一般会拟合得很好。在标准化赋值的情况下,β 相当于 $\rho/(1-\rho^2)$,其中 ρ 是隐含的连续变量之间的相关系数。当关联较弱时,$\beta\approx\rho$(参见:Becker 1989b;Goodman 1981a,b,1985)。

9.4.2 $L\times L$ 模型对应的相邻类别 Logit 模型

关于 $L\times L$ 模型的 Logit 表达式将 Y 视作结果变量,X 作为解释变量。令 $\pi_{j|i}=P(Y=j|X=i)$。结果变量的相邻类别 Logit(第 7.4.1 节)可表示为:

$$\log\frac{\pi_{j+1|i}}{\pi_{j|i}} = \log\frac{\mu_{i,j+1}}{\mu_{ij}} = (\lambda_{j+1}^Y - \lambda_j^Y) + \beta(v_{j+1} - v_j)u_i。$$

在 $\{v_j\}$ 间距相等的情况下,它简化为

$$\log\frac{\pi_{j+1|i}}{\pi_{j|i}} = \alpha_j + \beta u_i,$$

其中 $\alpha_j=\lambda_{j+1}^Y-\lambda_j^Y$。同一个线性 Logit 效应 β 适用于结果变量的所有 $(J-1)$ 对相邻类别:X 每变动一个单位,$Y=j+1$ 而不是 $Y=j$ 的发生比变动 e^β 倍。当对结果变量使用等距赋值时,我们隐含地假定对于 Y 的 $J-1$ 个相邻类别 Logit,X 的效应是相同的。

9.4.3 似然方程与模型拟合

对于式 9.6$L\times L$ 模型,泊松分布对数似然函数 $L(\boldsymbol{\mu}) = \sum_i\sum_j n_{ij}\log\mu_{ij} - \sum_i\sum_j\mu_{ij}$ 可简化为

$$L(\boldsymbol{\mu}) = n\lambda + \sum_i n_{i+}\lambda_i^X + \sum_j n_{+j}\lambda_j^Y + \beta\sum_i\sum_j u_i v_j n_{ij} - \sum_i\sum_j \exp(\lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j)。$$

将 $L(\boldsymbol{\mu})$ 对 $(\lambda_i^X,\lambda_j^Y,\beta)$ 求导,并令这三个偏导数等于零,便得到似然方程:

$$\hat{\mu}_{i+} = n_{i+},\quad i=1,\cdots,I,\qquad \hat{\mu}_{+j} = n_{+j},\quad j=1,\cdots,J,$$

$$\sum_i\sum_j u_i v_j \hat{\mu}_{ij} = \sum_i\sum_j u_i v_j n_{ij}。$$

利用如 Newton-Raphson 等迭代法可以得出最大似然估计。

令 $p_{ij}=n_{ij}/n$,$\hat{\pi}_{ij}=\hat{\mu}_{ij}/n$,以上第三个似然方程意味着

$$\sum_i\sum_j u_i v_j \hat{\pi}_{ij} = \sum_i\sum_j u_i v_j p_{ij}。$$

就拟合的和观察到的分布来说,由于边际分布,进而边际均值和方差都是一致的,第三个方程意味着,X 和 Y 的赋值之间的相关系数在两个分布中也是一样的。拟合计数表现出与观察数据相同的正向或负向趋势。

由于 $\{u_i\}$ 和 $\{v_j\}$ 是固定的,式 9.6$L\times L$ 模型仅比独立性模型多一个参数 (β)。它的残差自由度为

$$\text{df} = IJ - [1 + (I-1) + (J-1) + 1] = IJ - I - J,$$

除了 2×2 表格的情况外,它是一个非饱和模型。

9.4.4 例子:性行为的态度

表9.3也给出了双线性关联模型对该数据的拟合结果,其中行和列的赋值均为{1,2,3,4}。相应地,表9.4给出了软件的输出结果。为了拟合这个模型,我们在独立性模型中加入一个变量(由“linlin”表示),它的值等于行序号和列序号的乘积。与独立性模型($G^2(I)=127.6$, df=9)相比,$L\times L$ 模型的拟合结果大为改善[$G^2(L\times L)=11.5$, df=8]。对于独立性模型的预测中存在最大偏差的表格角部的单元格,拟合的改善尤其明显。

表9.4 对表9.3数据拟合双线性关联模型的输出结果

	Criteria For Assessing Goodness of Fit						
	Criterion		DF	Value			
	Deviance		8	11.533 7			
	Pearson Chi-Square		8	11.508 5			
			Standard	Wald 95%	Conf.	Chi-	
Parameter		Estimate	Error	Limits		Square	Pr > ChiSq
Intercept		0.473 5	0.433 9	-0.376 9	1.323 9	1.19	0.275 1
premar	1	1.753 7	0.234 3	1.294 4	2.212 9	56.01	<.000 1
premar	2	0.107 7	0.198 8	-0.282 0	0.497 4	0.29	0.588 0
premar	3	-0.016 3	0.126 4	-0.264 1	0.231 4	0.02	0.897 2
premar	4	0.000 0	0.000 0	0.000 0	0.000 0	.	.
birth	1	1.879 7	0.249 1	1.391 4	2.367 9	56.94	<.000 1
birth	2	1.415 6	0.199 6	1.024 3	1.806 8	50.29	<.000 1
birth	3	1.155 1	0.129 1	0.902 1	1.408 2	80.07	<.000 1
birth	4	0.000 0	0.000 0	0.000 0	0.000 0	.	.
linlin		0.285 8	0.028 2	0.230 5	0.341 2	102.46	<.000 1
			LR Statistics				
		Source	DF	Chi-Square		Pr > ChiSq	
		linlin	1	116.12		>.000 1	

译者注—premar:婚前性行为;birth:避孕服务;linlin:双线性项。

最大似然估计 $\hat{\beta}=0.286$(SE=0.028)表明,倾向于支持对青少年提供避孕服务的对象对婚前性行为也更宽容。模型估计的局部发生比之比等于 $\exp(\hat{\beta})=\exp(0.286)=1.33$。它的95%的沃尔德置信区间为 $\exp(0.286\pm1.96\times0.028)$,即(1.26, 1.41)。这个关联似乎并不是很强。但是,根据式9.7,非局部发生比之比的值要大得多。对表格四个角部的单元格所估计的发生比之比等于

$$\exp[\hat{\beta}(u_4-u_1)(v_4-v_1)]=\exp[0.286(4-1)(4-1)]=13.1。$$

这也可由关于角部单元格的拟合值计算得出,即$(80.9\times155.5)/(29.1\times33.0)=13.1$。

运用具有相同间距的两组赋值得出的 $\hat{\beta}$ 及拟合结果都相同。就上面的例子而言,其他任意一组等距赋值都会给出相同的拟合结果,但 $\hat{\beta}$ 的刻度相应发生变化。例如,利用行赋值为{2,4,6,8}和列赋值$\{v_j=j\}$也得到 $G^2=11.5$,但是 $\hat{\beta}=0.143$,SE=0.014(都大约为上面的一半)。关于表9.3中的数据,有人可能会认为第2类与第3类之间的距离比第1类和第2类或者第3类和第4类之间的距离大。例如,对行和列采用{1,2,4,5}的赋值就反映了这种想法。相应的 $L\times L$ 模型具有 $G^2=8.8$(df=8),以及 $\hat{\beta}=0.146$(SE=0.014)。

没有必要认为赋值必须反映类别之间的近似距离或者代表对定序变量的一种合理的测量，模型才会有效。它们仅仅表明在发生比之比中存在某种模式。如果运用等距的行赋值和列赋值的 $L \times L$ 模型拟合得很好，不论赋值是否代表了各类别之间的真实距离，一致性局部发生比之比都能描述这一关联。

对于表 9.3 中的赋值 $\{u_i = i\}$，婚前性行为所对应的边际均值和标准差分别为 2.81 和 1.26，相应的标准化赋值等于 $\{(i-2.81)/1.26\}$ 或 $(-1.44, -0.65, 0.15, 0.95)$。关于避孕服务的标准化等距赋值为 $(-1.65, -0.69, 0.27, 1.23)$。运用这些赋值，$\hat{\beta} = 0.374$。通过 $\hat{\beta} = \hat{\rho}/(1-\hat{\rho}^2)$ 对 $\hat{\rho}$ 求解，$\hat{\rho} = 0.333$。如果存在一个潜在的二元正态分布，我们估计两个变量之间的相关系数为 0.333。

9.4.5 定向的独立性定序检验

在双线性关联模型中，H_0：独立性等价于 $H_0: \beta = 0$。相应的似然比检验统计量等于

$$G^2(I \mid L \times L) = G^2(I) - G^2(L \times L)。$$

以发现正向或负向趋势为目的，该检验具有 $df = 1$。对于表 9.3 的数据，$G^2(I \mid L \times L) = 127.6 - 11.5 = 116.1$。这个检验的结果为 $P < 0.0001$，以非常明确的证据表明存在关联。相应的沃尔德统计量 $z^2 = (\hat{\beta}/SE)^2 = (0.286/0.0282)^2 = 102.5$ ($df = 1$) 也给出了同样的结果。在第 3.4.1 节中介绍的关于独立性检验的相关系数统计量公式 3.15 相当于模型中检验 $H_0: \beta = 0$ 的计分统计量，它等于 112.6 ($df = 1$)。

当 $L \times L$ 模型成立时，利用 $G^2(I \mid L \times L)$ 进行的定序检验比利用 $G^2(I)$ 的检验在渐近意义上更具效能。其原因与第 6.4.2 节中关于线性 Logit 模型的解释一致。对于给定的非中心参数，当自由度下降时，卡方检验的效能提高。在 $L \times L$ 模型成立的情况下，$G^2(I \mid L \times L)$ 和 $G^2(I)$ 的非中心参数是一样的；$G^2(I \mid L \times L)$ 的自由度为 $df = 1$ 而 $G^2(I)$ 的自由度为 $(I-1)(J-1)$，因而 $G^2(I \mid L \times L)$ 的统计效能更高。随着 I 和 J 的增加，这种效能优势会变得更明显。这是因为对 $G^2(I \mid L \times L)$ 而言，它的自由度仍是 $df = 1$，但 $G^2(I)$ 的自由度却会上升。

9.5 关联模型*

可以将双线性关联模型扩展到多维表格以及赋值本身作为参数而不是固定值的情况。这样的模型被称为*关联模型*（*association models*），因为它们主要关注变量之间的关联结构。

9.5.1 行效应和列效应模型

首先，我们介绍一个将 X 当作定类变量而将 Y 作为定序变量的模型。它适用于二维表格中各列之间具有排序的情况，将其赋值为 $v_1 \leqslant v_2 \leqslant \cdots \leqslant v_J$。由于各行之间是无序的，不需要对它们进行赋值。利用无序的参数 $\{\mu_i\}$ 替代式 9.6 模型中双线性项 $\beta u_i v_j$ 里的 $\{\beta u_i\}$，得到

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \mu_i v_j。 \tag{9.8}$$

它要求设定诸如 $\lambda_I^X = \lambda_J^Y = \mu_I = 0$ 的约束条件。这里的 $\{\mu_i\}$ 被称为*行效应*（*row effects*）。上述模型被称为*行效应模型*（*row effects model*）。

式 9.8 模型比独立性模型多包括 $I-1$ 个参数（这些 $\{\mu_i\}$）。独立性模型相当于 $\mu_1 =$

$\cdots=\mu_I$ 的一个特例。相应的*列效应模型*(*column effects model*)具有关联项 $u_i v_j$。它将 X 当作赋值为 $\{u_i\}$ 的定序变量,而 Y 作为具有参数 $\{v_j\}$ 的定类变量。行效应和列效应模型由 Goodman(1979a)、Haberman(1974b)以及 Simon(1974)提出并发展而来。

9.5.2 结果变量的相邻类别 Logit 模型

在 $\{v_{j+1}-v_j=1\}$ 的情况下,行效应模型具有相邻类别 Logit 的形式:

$$\log\frac{P(Y=j+1\mid X=i)}{P(Y=j\mid X=i)}=\alpha_j+\mu_i。\tag{9.9}$$

对于结果变量的每一对相邻类别,第 i 行的效应都相等。在 j 的不同取值下,将这些 Logit 按照 $i(i=1,\cdots,I)$ 进行绘图得到一组平行的线。Goodman(1983)将式 9.9 模型称为*平行发生比*(*parallel odds*)模型。

$\{\mu_i\}$ 之间的差异用于比较各行中关于 Y 的条件分布。当 $\mu_i=\mu_h$ 时,第 h 行和第 i 行具有相同的条件分布。如果 $\mu_i>\mu_h$,第 i 行中的 Y 则随机高于(stochastically higher)第 h 行中的 Y。

式 9.8 行效应模型的似然方程为 $\{\hat{\mu}_{i+}=n_{i+}\}$、$\{\hat{\mu}_{+j}=n_{+j}\}$,以及

$$\sum_j v_j\hat{\mu}_{ij}=\sum v_j n_{ij},\quad i=1,\cdots,I。$$

令 $\hat{\pi}_{j|i}=\hat{\mu}_{ij}/\hat{\mu}_{i+}$ 且 $p_{j|i}=n_{ij}/n_{i+}$。由于 $\hat{\mu}_{i+}=n_{i+}$,第三个似然方程可表示为 $\sum_j v_j\hat{\pi}_{j|i}=\sum_j v_j p_{j|i}$。对于每行中的条件分布,所拟合的列赋值的平均值等于样本分布的平均值。可以通过迭代法求解这些似然方程。

9.5.3 例子:政治观点

表 9.5 所示为威斯康辛州的一个选民样本在总统初选中政治观点与所属党派之间的关系。该表还给出了独立性模型(I)以及具有 $\{v_j=j\}$ 的行效应模型(R)的拟合值。

表 9.5 政治观点数据的观察频数与拟合值

党派	政治观点[a]			
	激进	中庸	保守	小计
民主党	143	156	100	399
	(102.0)[1]	(161.4)	(135.6)	
	(136.6)[2]	(168.7)	(93.6)	
独立党派	119	210	141	470
	(120.2)	(190.1)	(159.7)	
	(123.8)	(200.4)	(145.8)	
共和党	15	72	127	214
	(54.7)	(86.6)	(72.7)	
	(16.6)	(68.9)	(128.6)	

a 1 独立性模型;2 行效应模型。

来源:数据取自:R. D. Hedlund, *Public Opinion Quart.* 41:498-514(1978)。

表 9.6 给出了模型的输出结果。拟合优度检验表明,独立性模型是不充分的。加入行效应参数后,拟合结果大为改善($G^2(I)=105.7$,df=4;$G^2(R)=2.8$,df=2)。同时,利用 $G^2(I|R)=102.9$(df=2)对 $H_0:\mu_1=\mu_2=\mu_3$ 的检验也给出了存在关联的强有力证据。在表 9.5 中,拟合的改善在定序尺度的两端尤其明显,这里也是独立性模型出现偏差最大的地方。

表 9.6 对表 9.5 数据拟合行效应模型的输出结果

Criteria For Assessing Goodness of Fit

Criterion	DF	Value
Deviance	2	2.8149
Pearson Chi-Square	2	2.8039

Parameter		Estimate	Std Error	Wald 95% Conf. Limits		Chi Square	Pr > Chisq
Intercept		4.8565	0.0858	4.6883	5.0246	3204.02	<.0001
party	Pemoc	3.3230	0.3188	2.6981	3.9479	108.63	<.0001
party	Indep	2.9536	0.3149	2.3364	3.5707	87.98	<.0001
party	Repub	0.0000	0.0000	0.0000	0.0000	.	.
ideology	1	-2.0488	0.2216	-2.4831	-1.6145	85.50	<.0001
ideology	2	-0.6244	0.1139	-0.8476	-0.4013	30.08	<.0001
ideology	3	0.0000	0.0000	0.0000	0.0000	.	.
score * party	Democ	-1.2134	0.1304	-1.4690	-0.9577	86.56	<.0001
score * party	Indep	-0.9426	0.1260	-1.1896	-0.6956	55.95	<.0001
score * party	Repub	0.0000	0.0000	0.0000	0.0000	.	.

LR Statistics

Source	DF	Chi-Square	Pr > ChiSq
score * party	2	102.85	<.0001

译者注—Party：党派；Democ：民主党；Indep：独立党派；Repub：共和党；Ideology：政治观点；Score：赋值。

输出结果中，对每个变量的前两个类别使用了虚拟变量。交互项等于政治观点的赋值与党派的参数之间的乘积。因此，行效应的估计满足 $\hat{\mu}_3=0$，另外两个估计值分别表示前两个党派与共和党的对比。这些估计值分别为 $\hat{\mu}_1=-1.213$ 和 $\hat{\mu}_2=-0.943$。相对于共和党而言，$\hat{\mu}_i$ 的负值的绝对值越大，党派 i 的成员的政治观点也就越激进。在这个样本中，共和党员比其他两个党派都保守得多，而民主党员（第 1 行）是最激进的。根据式（9.9），模型预测在政治观点的相邻类别间发生比之比为常数。例如，由于 $\hat{\mu}_3-\hat{\mu}_1=1.213$，所估计的关于共和党政治观点保守而不是中庸，或中庸而不是激进的发生比是民主党的相应发生比的 exp(1.213) = 3.36 倍。图 9.3 显示了行效应模型中所估计的 Logit 的平行性特征。

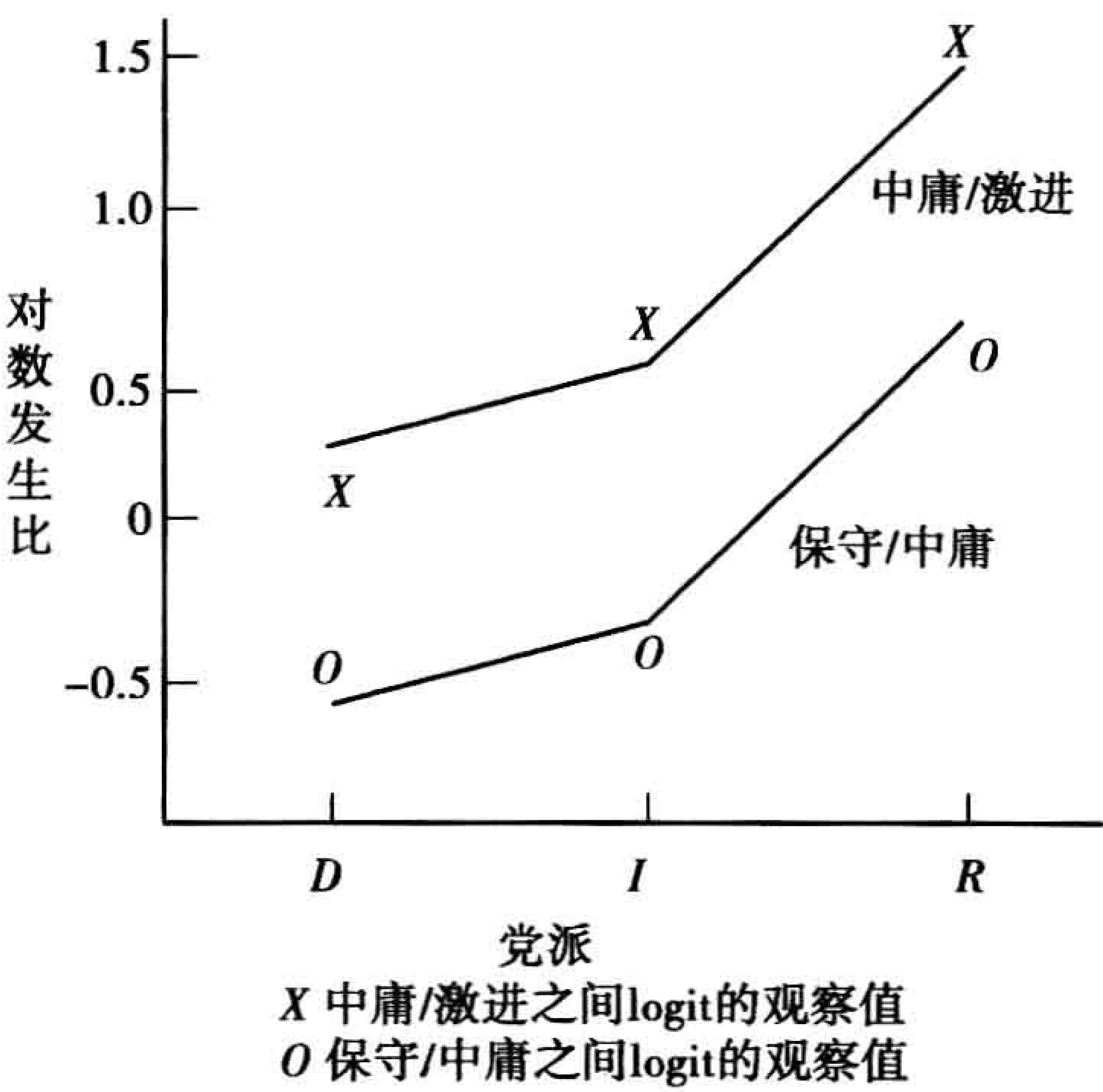


图 9.3 结果变量相邻类别 Logit 的观察值和预测值

对数线性模型本身并不区分结果变量和解释变量。相应地，大家也可以利用累积 Logit 模型来描述党派对政治观点的效应，或者利用基线类别 Logit 模型来描述政治观点对党派的线性效应。

9.5.4 多维表格模型中的定序变量

对关联模型进行的扩展可以用来分析多维表格中的定序变量。在三维表格的情况下，有很多模型可供选择：①比定类模型（XY, XZ, YZ）更简约的关联模型；②允许异质性关联但又不同于饱和模型（XYZ）的模型。

比模型（XY, XZ, YZ）更简约的关联模型，用一个反映排序特征的结构项来替代 λ 关联项。例如，当 X 和 Y 都是定序变量时，有关 λ_{ij}^{XY} 的其他选择包括双线性项 $\beta u_i v_j$、行效应项 $\mu_i v_j$ 或者列效应项 $u_i v_j$；这些选项分别设定了一种在行和列内、只在行内或只在列内的条件分布的随机排序。在包括双线性项的情况下，模型可表示为

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}。 \quad (9.10)$$

这时，对于所有的 k，条件局部发生比之比式 8.13 满足：

$$\log \theta_{ij(k)} = \beta(u_{i+1} - u_i)(v_{j+1} - v_j)。$$

在不同的分表中该关联都相同，因而称之为*同质双线性 XY 关联*（*homogeneous linear-by-linearXY association*）。

当存在异质性关联时，与饱和模型相比，关于定序变量的结构项使得模型参数更易于解释。例如，*异质双线性 XY 关联模型*（*heterogeneous linear-by-linearXY association model*）

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_k u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad (9.11)$$

允许 XY 关联随着 Z 的不同取值而变化。在使用单位距离的赋值时，

$$对所有的\ i\ 和\ j,\ \log \theta_{ij(k)} = \beta_k。$$

在 Z 的每个取值内，它存在一致性关联；但是在 Z 的不同取值之间，关联的强度存在差异。拟合该模型相当于在 Z 的每个取值上分别拟合式 9.6$L \times L$ 模型。

9.5.5 例子：空气污染与呼吸检查结果

表 9.7 展示了在德克萨斯州休斯顿的一些工厂中工人的吸烟状况（S）、呼吸检查结果（B）以及年龄（A）之间的关联。对数线性模型（SA, SB, BA）拟合得很差（$G^2 = 25.9$，df = 4）。因而，更简单的模型如同质双线性 SB 关联模型是不可行的（在使用等距赋值的情况下，$G^2 = 29.1$，df = 7）。在模型中仅增加一个参数后的异质双线性 SB 关联模型拟合

表 9.7 按照呼吸检查结果对行业工人的交叉分组

年　龄	吸烟状况	呼吸检查结果		
		正常	不确定	不正常
<40	从未吸烟	577	27	7
	曾吸烟	192	20	3
	现在吸烟	682	46	11
40 ~ 59	从未吸烟	164	4	0
	曾吸烟	145	15	7
	现在吸烟	245	47	27

来源：From p. 21 of *Public Program Analysis* by R. N. Forthofer and R. G. Lehnen. Copyright ©1981 by Lifetime Learning Publications, Belmont, CA 94002, a division of Wadsworth, Inc. 经 Van Nostrand Reinhold 授权重印。保留所有权利。

得很好($G^2=10.8$,df = 6)。在对 S 和 B 使用整数赋值的情况下,年龄较小一组对应于 $\hat{\beta}_1=0.115$,年龄较大一组对应于 $\hat{\beta}_2=0.781$,二者之差的标准误为 SE = 0.167。对于较年长一组来说,吸烟的效应看起来要强很多,该组中所估计的局部发生比之比等于 exp(0.781) = 2.18,而对较年轻一组,相应发生比之比仅为 exp(0.115) = 1.12。对该数据,可能以 B 为结果变量拟合 Logit 模型更适当一些(习题 7.11)。

当层变量是定序变量时,在各层间某些对数发生比之比可能存在线性趋势,如表 9.8 所示。该数据来自于一个煤矿工人的样本,收集的变量包括 B = 无法呼吸、W = 哮喘以及 A = 年龄,其中 B 和 W 是结果变量。我们可以分别通过 Logit 模型来描述年龄对每个结果变量的效应。为了分析 BW 关联是否随着年龄变动,我们拟合模型(BW,AB,AW)。它的拟合结果具有 $G^2=26.7$,df = 8。表 9.8 给出了模型的标准化皮尔逊残差。随着年龄的上升,这些残差显示出下降的趋势。

表 9.8 按照无法呼吸、哮喘以及年龄划分的煤矿工人

年龄	无法呼吸				标准化皮尔逊残差[a]
	是		否		
	哮喘 是	哮喘 否	哮喘 是	哮喘 否	
20 ~ 24	9	7	95	1 841	0.75
25 ~ 29	23	9	105	1 654	2.20
30 ~ 34	54	19	177	1 863	2.10
35 ~ 39	121	48	257	2 357	1.77
40 ~ 44	169	54	273	1 778	1.13
45 ~ 49	269	88	324	1 712	-0.42
50 ~ 54	404	117	245	1 324	0.81
55 ~ 59	406	152	225	967	-3.65
60 ~ 64	372	106	132	526	-1.44

a 残差对应于"是-是"和"否-否"单元格;"是-否"和"否-是"单元格的残差符号与此相反。

来源:授权重印自:Ashford and Sowden (1970)。

上述结果表明,应当考虑模型

$$\log \mu_{ijk} = (BW,AB,AW) + kI(i=j=1)\delta, \tag{9.12}$$

其中,I 是一个指示函数(indicator function)。该模型在同质性关联的基础上,对单元格 μ_{111} 加上 δ,…,对单元格 μ_{119} 加上 9δ。这时,关于 BW 的对数发生比之比在各个年龄类别之间呈线性变动。模型拟合结果为 $\hat{\delta}=-0.131$(SE = 0.029)。在年龄组取值为 k 时,所估计的 BW 的对数发生比之比等于 $3.676-0.131k$,其变动范围为 3.55 至 2.50。该模型的拟合结果具有 $G^2=6.80$(df = 7)。McCullagh 和 Nelder(1989, sec. 6.6)对此数据进行了其他分析。

9.5.6 关于条件独立性的其他定序检验

可以通过对 $G^2(I|L\times L)$ 的扩展,进行定序尺度下的条件独立性检验,例如,对比 XY 条件独立性模型(XZ,YZ)与同质双线性 XY 关联模型式 9.10,它相当于在后一模型中检验 $\beta=0$,相应的 df = 1。这是第 7.5.3 节介绍的条件独立性定序检验的另一种方法。与 Mantel 计分检验式 7.21 一样,由于 $\sum_k(\sum_i\sum_j u_i v_j n_{ijk})$ 是式 9.10 模型中 β 的充分统计

量，这个统计量使用了相关系数的信息。事实上，Mantel 统计量相当于在该模型中检验 $H_0: \beta = 0$ 的计分统计量。

使用这些模型的似然比、计分或沃尔德统计量可以进行小样本下的精确检验。进行这些检验需要用到特殊的算法（Agresti et al. 1990；Kim and Agresti，1997）。

9.6 关联模型、相关模型与对应分析*

双线性关联（$L \times L$）模型是行赋值为参数的行效应（R）模型以及列赋值为参数的列效应（C）模型的一个特例。这些模型又都是行赋值和列赋值都为参数的更一般化的模型的特例。

9.6.1 可积行效应和列效应模型

在式 9.6$L \times L$ 模型中，用参数替代 $\{u_i\}$ 和 $\{v_j\}$ 就得到了*行列效应*（*row and column effects*）（RC）模型（Goodman，1979a）

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta\mu_i v_j。\qquad (9.13)$$

模型的可识别性要求对 $\{\mu_i\}$ 和 $\{v_j\}$ 进行位置（location）和规模（scale）的限定。它的残差自由度为 $df = (I-2)(J-2)$。该模型*不属于*对数线性模型，因为它的预测项包括参数 μ_i 和 v_j 之间的可积（而不是线性）函数。该模型将变量视为定类变量，行或列类别次序的变动不会影响拟合结果。当至少有一个变量是定序变量时，存在对参数的简单解释，即可以将其解释为局部对数发生比之比

$$\log \theta_{ij} = \beta(\mu_{i+1} - \mu_i)(v_{j+1} - v_j)。$$

尽管使用参数而不是随意的赋值似乎更具吸引力，但是 RC 模型会导致一些对数线性模型不会遇到的复杂问题。它的似然函数可能不是凹函数，并且可能会存在局部最大值（local maxima）。虽然独立性模型是它的一个特例，但是通过 RC 模型来进行独立性检验却并不容易。Haberman（1981）表明，$G^2(I) - G^2(RC)$ 不服从卡方零分布，却服从一个具有 Wishart 矩阵的最大特征值的分布。

当其中一组参数是固定的赋值时，RC 模型简化为 R 模型或 C 模型。利用这个特点，Goodman（1979a）提出了一种迭代模型拟合算法来拟合 RC 模型。该算法的每个循环包括两个步骤。首先，对 $\{v_j\}$ 的某些初始猜测值，如同在 R 模型中那样，去估计行的赋值。接着，将在第一步中所估计的行赋值视为固定的，再像在 C 模型中那样去估计列的赋值。这些估计值又作为下一个循环中第一步的固定列赋值，以重新估计 R 模型中的行赋值。这种算法无法保证模型会收敛于最大似然估计，但当模型拟合得很好时，它一般会收敛。Haberman（1995）介绍了拟合关联模型的更复杂的方法。

Goodman（1985）将饱和模型中的关联项表示为一种对 RC 模型中的 $\beta\mu_i v_j$ 项的扩展形式，即

$$\lambda_{ij}^{XY} = \sum_{k=1}^{M} \beta_k \mu_{ik} v_{jk}, \qquad (9.14)$$

其中 $M = \min(I-1, J-1)$。模型参数满足如下约束条件：

$$\text{对所有 } k,\quad \sum_i \mu_{ik}\pi_{i+} = \sum_j v_{jk}\pi_{+j} = 0,$$

$$\text{对所有 } k,\quad \sum_i \mu_{ik}^2\pi_{i+} = \sum_j v_{jk}^2\pi_{+j} = 1, \qquad (9.15)$$

$$\text{对所有 } k \neq h,\ \sum_i \mu_{ik}\mu_{ih}\pi_{i+} = \sum_j v_{jk}v_{jh}\pi_{+j} = 0。$$

在 $k > M^*$ 的情况下，当式 9.14 模型中 $\beta_k = 0$ 时，它被称为 $RC(M^*)$ 模型。对此模型的最大似然拟合，参见：Becker(1990)。式 9.13*RC* 模型是其中 $M^* = 1$ 的特例。

9.6.2　例子：精神健康状况

表 9.9 描述了曼哈顿居民某一样本中童年精神健康状况与父母的社会经济地位之间的关系(Goodman, 1979a)。*RC* 模型对该数据拟合得很好($G^2 = 3.6$, df = 8)。利用公式 9.15 的参数约束条件，关于行赋值的最大似然估计为(-1.11, -1.12, -0.37, 0.03, 1.01, 1.82)，关于列赋值的最大似然估计为(-1.68, -0.14, 0.14, 1.41)，并且 $\hat{\beta} = 0.17$。几乎所有局部对数发生比之比的估计值都是正的，这表明存在父母社会经济地位越高、孩子的精神健康状况也越好的趋势。

此外，定序对数线性模型也拟合得很好。在使用等距赋值的情况下，$G^2(L \times L) = 9.9$ (df = 14)。统计量 $G^2(L \times L | RC) = 6.3$ (df = 6) 用于检验 *RC* 模型中的行赋值和列赋值是否是等距的。参数形式的赋值并没能显著提高模型的拟合结果。因此，利用一致性局部发生比之比来描述该表格的关联是充分的。在使用单位距离的赋值的情况下，$\hat{\beta} = 0.091$ (SE = 0.015)，拟合的局部发生比之比等于 exp(0.091) = 1.09。这给出了存在正向关联的强有力证据，不过，至少就局部发生比之比而言，关联的强度较弱。

表 9.9　精神健康状况与社会经济地位的交叉分组

父母社会经济地位	精神健康状况			
	很好	轻微症状	中度症状	受损
A　高	64	94	58	46
B	57	94	54	40
C	57	105	65	60
D	72	141	77	94
E	36	97	54	78
F　低	21	71	54	71

来源：授权重印自：L. Srole et al. *Mental Health in the Metropolis: The Midtown Manhattan Study*, (New York: NYU, Press, 1978), p. 289。

9.6.3　相关模型

二维表格的*相关模型*(*correlation model*)在许多方面与 *RC* 模型相同(Goodman, 1985)。以最简单的形式为例，相关模型可表示为

$$\pi_{ij} = \pi_{i+}\pi_{+j}(1 + \lambda\mu_i v_j),\tag{9.16}$$

其中 $\{\mu_i\}$ 和 $\{v_j\}$ 是关于行、列赋值的参数，它们满足

$$\sum \mu_i \pi_{i+} = \sum v_j \pi_{+j} = 0 \quad \text{以及} \quad \sum \mu_i^2 \pi_{i+} = \sum v_j^2 \pi_{+j} = 1。$$

式 9.16 联合分布中的参数 λ 是两组赋值之间的相关系数。

相关模型也被称为*典型相关模型*(*canonical correlation model*)，这是因为关于赋值的最大似然估计使式 9.16 模型中的相关系数取最大值。典型相关模型的一般形式为

$$\pi_{ij} = \pi_{i+}\pi_{+j}\left(1 + \sum_{k=1}^{M} \lambda_k \mu_{ik} v_{jk}\right),$$

其中$0\leqslant\lambda_M\leqslant\cdots\leqslant\lambda_1\leqslant1$，它具有类似于式9.15的约束条件。参数$\lambda_k$是$\{\mu_{ik},i=1,\cdots,I\}$和$\{v_{jk},j=1,\cdots,J\}$之间的相关系数。$\{\mu_{i1}\}$和$\{v_{j1}\}$是使联合分布中的相关系数$\lambda_1$最大化的标准化赋值；$\{\mu_{i2}\}$和$\{v_{j2}\}$是使相关系数$\lambda_2$最大化的标准化赋值，并保证$\{\mu_{i1}\}$与$\{\mu_{i2}\}$不相关以及$\{v_{j1}\}$与$\{v_{j2}\}$不相关；以此类推。

非饱和模型等于用$M^*<\min(I-1,J-1)$替代上述模型中的M。Gilula和Haberman(1986)以及Goodman(1985)讨论了对相关模型的最大似然拟合。Goodman(1981a,1985,1986)指出，当式9.16中的λ接近于零时，关于λ和赋值参数的最大似然估计与RC模型中对β和赋值参数的估计相似。在相关模型中，也可以使用固定的而不是参数形式的赋值。

Goodman探讨了关联模型相对于相关模型所具有的优势。由于存在约束条件$0\leqslant\pi_{ij}\leqslant1$，相关模型并不是对所有可能的赋值情况都成立；它的最大似然拟合值并不具有与观察数据相同的边际总计；并且，这个模型无法轻易地扩展到多维表格的情况。Gilula和Haberman(1988)通过将多维表格中的所有解释变量视为一个单一变量并将所有结果变量视为另一个变量，从而利用相关模型进行了分析。

9.6.4 对应分析

对应分析(correspondence analysis)是一种利用绘图来反映二维列联表中的关联关系的方法。行和列由图中的点所表示，用点的位置来表示关联。Goodman(1985,1986)指出，这些点的坐标相当于将一般典型相关模型中的$\{\mu_{ik}\}$和$\{v_{jk}\}$重新进行参数设定(reparameterizations)。对应分析使用的是调整后的赋值

$$x_{ik}=\lambda_k\mu_{ik},\quad y_{jk}=\lambda_k v_{jk}。$$

当第k个维度的相关系数λ_k接近于零时，这些坐标也接近于零。对应分析图利用了前两个维度，对每一行画出(x_{i1},x_{i2})，对每一列画出(y_{j1},y_{j2})。

Goodman(1985,1986)利用表9.9的数据，展示了对应分析与相关模型和关联模型之间的相似性。在一般化的典型相关模型中，$M=\min(I-1,J-1)=3$。它所估计的相关系数的平方为(0.026 0, 0.001 4, 0.000 3)。由此可见，这个关联相当弱。表9.10给出了关于这三个维度的对应分析所估计的行赋值和列赋值。除了前两个行赋值略有不同外，第一个维度上的两组赋值都存在单调递增的模式。第二个和第三个维度的赋值都接近于零，反映出$\hat{\lambda}_2$和$\hat{\lambda}_3$相对较小。

表9.10 对表9.9数据进行对应分析所得到的赋值

列赋值	维度			行赋值	维度		
	1	2	3		1	2	3
1	0.260	0.012	0.023	1	0.181	0.018	0.028
2	0.030	0.024	-0.019	2	0.185	-0.011	-0.026
3	-0.013	-0.069	-0.002	3	0.059	-0.021	-0.010
4	-0.236	0.019	0.016	4	-0.008	0.042	0.011
				5	-0.164	0.044	-0.009
				6	-0.287	-0.061	0.005

来源：授权重印自：the Institute of Mathematical Statistics, based on Goodman(1985)。

图9.4显示了对应分析的结果。水平轴是关于第一个维度的估计，垂直轴是关于第二个维度的估计。六个圆形的点代表了六行，其中第i个点给出了$(\hat{x}_{i1},\hat{x}_{i2})$。相似地，四

个方形的点显示了估计值($\hat{y}_{j1}$,$\hat{y}_{j2}$)。这两组点都落在与水平轴相近的地方,因为第一个维度比第二个维度更重要。

图 9.4 中,代表行的点都比较接近,这表明在各列之间关于行的条件分布相似。代表列的点相近则表明在各行之间关于列的条件分布相似。代表行的点与代表列的点相近反映这些组合值发生的可能性比独立性假设下所预期的大。图 9.4 显示出当一个变量取值较大时另一个变量取值也较大、一个变量取值较小时另一个变量取值也较小的趋势。

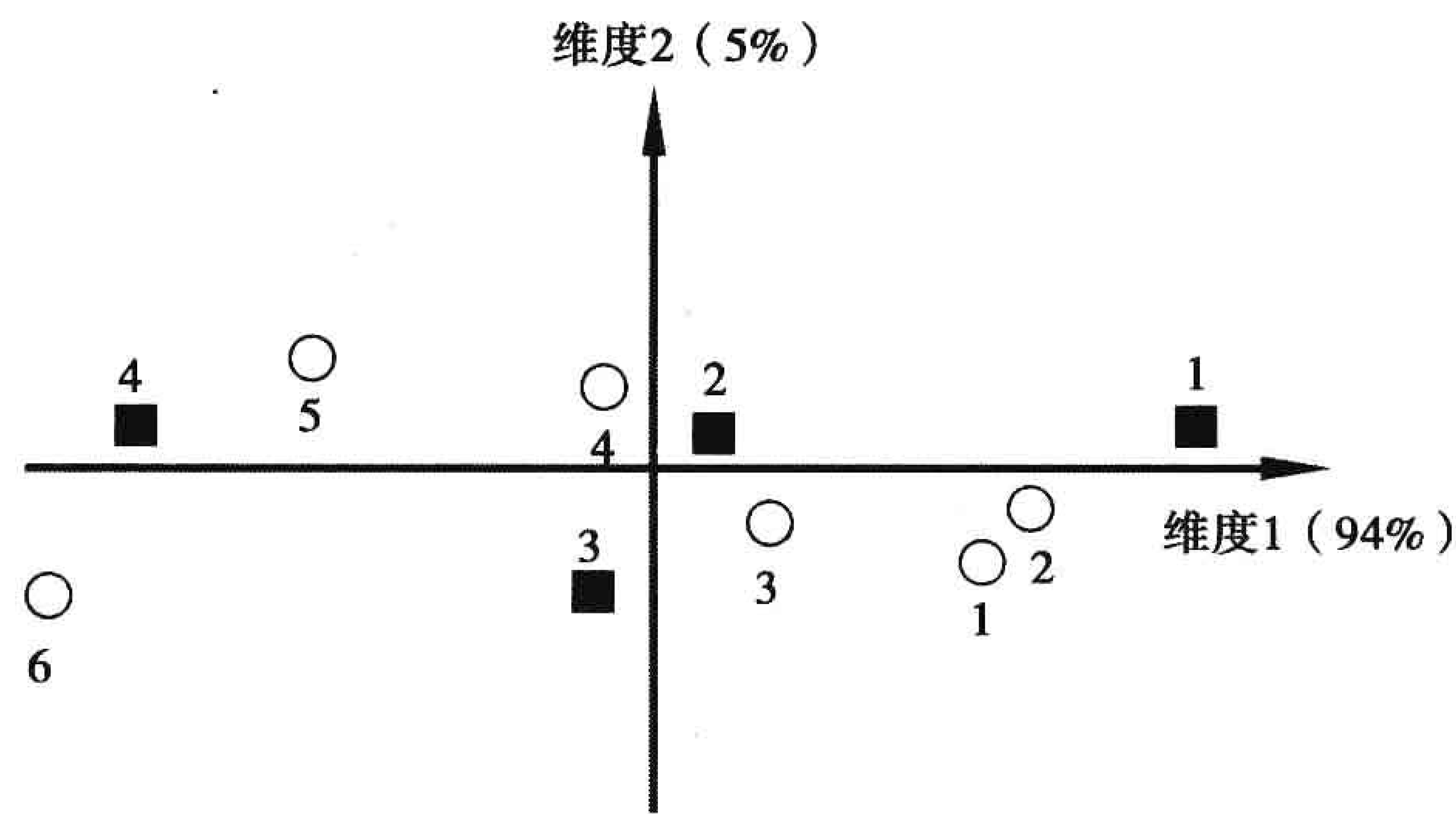


图 9.4　对应分析中前两个维度的赋值(授权重印自:Escoufier(1982))

对应分析主要被用作一种描述性的工具。Goodman(1986)发展了有关的统计推断方法。对于表 9.9,推断分析发现,第一个维度占所有相关系数的平方的 94%,它能够充分地描述该关联。Goodman 推荐使用仅包括一个维度且只显示该维度的拟合赋值的非饱和模型。这时,对应分析等价于利用式 9.16 相关模型进行的最大似然分析。这个模型所估计的行赋值为(-1.09,-1.17,-0.37,0.05,1.01,1.80),列赋值为(-1.60,-0.19,0.09,1.48)。模型拟合得很好(G^2=2.75,df=8)。拟合优度和所估计的赋值与在第 9.6.2 节中 *RC* 模型的结果相似。更简约的相关模型也能很好地拟合这些数据,比如使用等距赋值的模型。

对表 9.9 的数据进行的所有分析都给出了相似的结论。但是,它们都忽略了精神健康实际上是一个自然的结果变量。在这里,使用定序 Logit 模型可能更合理一些。

与相关模型相似,对应分析的一个严重缺陷在于,它无法直接扩展到多维表格的情况。Greenacre(1993)在一张图中显示了多个成对的关联。

9.6.5　关于定序变量的模型选择与赋值选择

以上三节介绍了几种在模型构建过程中应用变量类别排序特征的方法。在允许定序效应的情况下,可考虑的潜在模型远远多于标准的对数线性模型。为了选择适当的模型,一种可行的方法是利用标准的模型作为指导。如果一个标准的模型拟合得很好,通过将其中的某些参数替代为定序的结构项可以对其进行简化。

关联模型、相关模型和对应分析都需要对定序变量的类别进行赋值。在等距赋值的情况下,对参数的解释是最简单的。在将赋值作为参数时,所得到的关于赋值的最大似然估计不一定是单调的。通过对模型加以限定,对满足排序要求的似然函数进行最大化可以确保赋值之间具有单调性(如 Agresti et al. 1987;Ritov and Gilula, 1991)。但是,将赋值作为参数也存在缺点。由于需要的自由度较多,模型变得不那么简约,对效应进行检验的统计效能下降(回顾第 6.4.3 节)。当只有一个变量是结果变量时,对其拟合累积连结模型(第 7.2 和 7.3 节)不需要事先指定或使用参数形式的赋值。

9.7 关于比率的泊松回归

对数线性模型的应用并不一定局限于列联表数据。在第 4.3 节中，我们介绍了对计数数据进行分析的泊松回归模型。当结果发生在某个时间、空间或其他表示范围的指标内时，利用模型分析该结果的*发生率*(*rate*)比分析发生数更有意义。

9.7.1 利用包括抵消项的对数线性模型分析比率

当结果变量计数 n_i 的基数(index)为 t_i 时，样本比率为 n_i/t_i。它的期望值等于 μ_i/t_i。由 x 表示解释变量，关于比率的期望值的对数线性模型具有以下形式：

$$\log(\mu_i/t_i) = \alpha + \beta x_i。 \quad (9.17)$$

这个模型也可以等价地表示为

$$\log \mu_i - \log t_i = \alpha + \beta x_i。$$

如第 8.7.4 节所述，关于均值的对数连结的调整项，即 $-\log t_i$，被称为*抵消项*(*offset*)。模型的拟合相当于将 $\log t_i$ 作为模型右边的一个预测项，并设定它的系数等于 1.0。

在式 9.17 模型中，结果变量的期望计数满足

$$\mu_i = t_i \exp(\alpha + \beta x_i)。$$

均值与基数成比例，其中，有关比例的常数取决于 x 的取值。在使用恒等连结的情况下，上述模型可表示为：

$$\mu_i/t_i = \alpha + \beta x_i, \quad 或者 \quad \mu_i = \alpha t_i + \beta x_i t_i。$$

它不再需要抵消项。这个模型相当于使用恒等连结的普通泊松分布广义线性模型，以 t_i 和 $x_i t_i$ 为解释变量，而且模型不包括截距项。它给出的是预测变量之间的可加效应，而不是可积效应。当存在多个预测变量时，由于在迭代过程中可能会出现负的拟合计数而导致拟合过程无法继续，这使得使用恒等连结的模型的应用价值大为降低。

9.7.2 关于心脏瓣膜手术死亡率的模型分析

Laird 和 Olivier(1981)分析了病人在接受心脏瓣膜置换手术后的存活情况。该样本共包括 109 位病人，数据给出了心脏瓣膜的类型(动脉瓣，二尖瓣)以及患者年龄(<55，≥55)。该研究对病人进行了跟踪调查，直到病人死亡或研究期限截止。手术发生在整个研究期间，跟踪期从 3 个月到 97 个月不等。结果变量是该病人在研究截止时是否已经死亡以及跟踪期的长度。对于已经死亡的对象，这个时间就是从手术结束到死亡之间的时长；对其他对象，它等于从手术结束直到研究截止或研究对象退出研究的时间。

表 9.11 列出了按心脏瓣膜类型和年龄划分的跟踪期内所发生的死亡数。这些计数是关于心脏瓣膜类型、年龄以及是否死亡(是，否)的三维列联表的第一层。表 9.11 中未列出的对象在跟踪期结束时仍然活着。这些对象*被删失掉了*(*censored*)，因为我们只知道他们在术后存活时长的下限。利用关于死亡概率的二分广义线性模型来分析这个 2×2×2 表格是不适当的，因为各对象的历险时间不同；将仅观察了 3 个月和观察了 97 个月的对象作为具有相同概率的同一试验是不合理的。为了利用年龄和瓣膜类型作为预测变量对死亡频数进行分析，适当的基数不是研究对象的数量而是这些对象的历险总时长。因此，我们对死亡率进行模型分析。

一个对象的*历险时间*(*time at risk*)等于他的跟踪期的长度。对于给定的年龄和瓣膜类型，历险总时长等于相应单元格中所有对象的历险时间的总和(包括那些死亡的和删

失的对象)。表 9.11 列出了以月为单位的总时长。该表还给出了样本比率,即死亡数除以历险总时长。例如,对于进行了动脉瓣膜置换手术的年轻对象,在 1 259 个月的观察期内发生了 4 例死亡,所以,相应的样本比率等于 4/1 259 = 0.003 2。

表 9.11　关于心脏瓣膜置换手术的数据

年龄		心脏瓣膜类型	
		动脉瓣	二尖瓣
<55	死亡数量	4	1
	历险时间	1 259	2 082
	死亡率	0.003 2	0.000 5
55 +	死亡数量	7	9
	历险时间	1 417	1 647
	死亡率	0.004 9	0.005 5

来源:授权重印,数据基于:Laird and Olivier(1981)。

现在,我们通过模型来分析年龄和瓣膜类型对这个比率的效应。令 a 表示关于年龄的虚拟变量,其中 $a_1=0$ 表示年轻组,$a_2=1$ 表示年老组。令 v 表示关于瓣膜类型的虚拟变量,其中 $v_1=0$ 表示动脉瓣,$v_2=0$ 表示二尖瓣。令 n_{ij} 表示年龄组 a_i、瓣膜类型 v_j 所发生的死亡数量,它在整个历险时间 t_{ij} 内的期望值为 μ_{ij}。给定 t_{ij},期望比率等于 μ_{ij}/t_{ij}。模型

$$\log(\mu_{ij}/t_{ij}) = \alpha + \beta_1 a_i + \beta_2 v_j \tag{9.18}$$

假定年龄和瓣膜类型的效应之间不存在交互项。

可以利用标准的迭代法拟合该模型,将 $\{n_{ij}\}$ 作为均值为 $\{\mu_{ij}\}$ 的独立泊松分布变量。该模型拟合需要以 $\{t_{ij}\}$ 为条件。表 9.12 给出了所拟合的死亡计数和估计的死亡率。相应的效应估计值分别为

$$\hat{\beta}_1 = 1.221 \quad (SE = 0.514), \quad \hat{\beta}_2 = -0.330 \quad (SE = 0.438)。$$

这一结果表明,存在着显著的年龄效应。对于给定的瓣膜类型,年老组所估计的死亡率是年轻组的 exp(1.221) = 3.4 倍。关于 β_1 的 95% 的沃尔德置信区间为 1.221 ± 1.96(0.514),这可转化为关于真实的可积效应 exp(β_1)的置信区间(1.2,9.3)(相应的似然比置信区间为(1.3,10.4))。这项研究包含了相当多的删失数据。在 109 名病人中,只有 21 名在观察期内死亡。因而,上述模型对两个效应的估计都是不精确的。尽管如此,这个分析还是利用了所有 109 名病人对历险时间的贡献。

表 9.12　关于表 9.11 数据的泊松回归模型的拟合结果

年龄		对数连结		恒等连结	
		动脉瓣	二尖瓣	动脉瓣	二尖瓣
<55	死亡数量	2.28	2.72	3.16	1.19
	死亡率	0.001 8	0.001 3	0.002 5	0.000 6
55 +	死亡数量	8.72	7.28	9.17	7.48
	死亡率	0.006 2	0.004 4	0.006 5	0.004 6

将 $\{n_{ij}\}$ 与拟合值 $\{\hat{\mu}_{ij}\}$ 进行对比的拟合优度统计量分别为 $G^2=3.2$ 和 $X^2=3.1$。因为模型通过三个参数来拟合四个结果变量计数,残差自由度为 df = 1。模型表现出略微的拟合不足,这说明可能存在关于瓣膜类型和年龄之间的交互项。但是,剔除瓣膜类型效应的模型(即,令式 9.18 中的 $\beta_2=0$)拟合得几乎一样好,它的 $G^2=3.8$,$X^2=3.8$(df = 2)。

在去除掉年龄效应后,模型拟合得很差。

相应的使用恒等连结的模型为

$$\mu_{ij} = \alpha t_{ij} + \beta_1 a_i t_{ij} + \beta_2 v_j t_{ij},$$

它的拟合结果非常好,$G^2 = 1.1$,$X^2 = 1.1$(df = 1)。表9.12 显示了这个拟合结果。两个模型得出的实质性结论相似。在使用恒等连结的模型中,估计值 $\hat{\beta}_1 = 0.0040$ (SE = 0.0014)表示的是对于每种类型的心脏瓣膜,所估计的年轻组与年老组之间死亡率的差异。

9.7.3 关于生存时间的模型分析*

利用模型分析生存时间的方法与关于比率的泊松分布对数线性模型有关。这种方法关注的是死亡发生的时间,而不是死亡的数量。令 T 表示某个事件发生的时间,如死亡或者在可靠性研究(reliability study)中产品出现故障。令 $f(t)$ 和 $F(t)$ 分别表示 T 的概率密度函数(pdf)和累积分布函数(cdf)。利用泊松分布似然函数对事件数量的最大似然估计与利用负指数分布似然函数对 T 的估计之间存在一定联系(Aitkin and Clayton, 1980)。

对于 $T = t$ 的一个对象,其对似然函数的贡献为 $f(t)$。对于删失时间等于 t 的对象,我们仅知道 $T > t$,因此,其对似然函数的贡献为 $P(T > t) = 1 - F(t)$。针对第 i 个对象,利用 $w_i = 1$ 表示死亡,$w_i = 0$ 表示删失,那么关于 n 个独立观测值的生存时间似然函数为

$$\prod_{i=1}^{n} f(t_i)^{w_i}[1 - F(t_i)]^{1-w_i}。$$

相应的对数似然函数等于

$$\sum_i w_i \log[f(t_i)] + \sum_i (1 - w_i)\log[1 - F(t_i)]。\qquad (9.19)$$

进一步的分析要求给出关于 f 的参数形式以及使其参数取决于解释变量的模型。

大多数生存模型关注死亡所发生的率(*rate*)而不是 $E(T)$。*风险函数*(*hazard function*)

$$h(t) = \frac{f(t)}{1 - F(t)} = \lim_{\varepsilon \downarrow 0} \frac{P[t < T < t + \varepsilon \mid T > t]}{\varepsilon}$$

反映的是已经存活到时间 t 的对象的瞬时死亡率。关于生存模型的一个简单的密度函数是负指数分布函数。它的概率密度函数为

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0。$$

在 $t > 0$ 时,它的累积分布函数为 $F(t) = 1 - e^{-\lambda t}$,并且 $E(T) = \lambda^{-1}$。风险函数

$$h(t) = \lambda, \quad t > 0,$$

对所有的 t 都等于常数。

现在,我们引入解释变量 $\mathbf{x}$。假定负指数生存分布的风险函数为

$$h(t;\mathbf{x}) = \lambda \exp(\boldsymbol{\beta}'\mathbf{x})。\qquad (9.20)$$

也即,T 的分布中包括的参数通过式 9.20 取决于 $\mathbf{x}$。对于解释变量效应的函数形式式 9.20的选择,需要确保在 $\mathbf{x}$ 的所有取值上风险都非负。例如,式 9.18 对数线性模型对应着一个类似式 9.20 的关于比率的可积模型。

考虑式 9.19 对数似然函数中 $f(t)$ 等于参数为 $\lambda\exp(\boldsymbol{\beta}'\mathbf{x})$ 的负指数密度函数。对于第 i 个对象,令

$$\mu_i = t_i \lambda \exp(\boldsymbol{\beta}'\mathbf{x}_i)。$$

将其代入相应公式,对数似然函数简化为

$$\sum_i w_i \log \mu_i - \sum_i \mu_i - \sum_i w_i \log t_i。$$

上式的前两项与 $\boldsymbol{\beta}$ 有关。这部分与期望值为 $\{\mu_i\}$ 的独立泊松变量 $\{w_i\}$ 的对数似然函数一致。不同的是,在上式中 $\{w_i\}$ 是二分变量而不是泊松变量,但是这并不影响对 $\boldsymbol{\beta}$ 的最大化过程。这个过程等价于,利用观测值 $\{w_i\}$ 来最大化包括抵消项 $\log(t_i)$ 的泊松分布对数线性模型

$$\log \mu_i - \log t_i = \log \lambda + \boldsymbol{\beta}'\mathbf{x}_i$$

的似然函数。当我们将对数似然函数中所有 $\mathbf{x}$ 值相同的对象对应的项加总时,可以得到在 $\mathbf{x}$ 的每个取值对应的死亡数($\sum w_i$),同时抵消项等于每个取值处的($\sum t_i$)的对数。

多数情况下,关于风险不随时间的变动而变动的假定并不合理。随着产品的损耗,其故障发生率会上升。一种扩展方法是将时间划分为离散的区间,并假定在每个区间内风险恒定,即,对于第 k 个区间的 t,

$$h(t;\mathbf{x}) = \lambda_k \exp(\boldsymbol{\beta}'\mathbf{x}),$$

其中 $k=1,\cdots$。这样,在每个时间区间内都适用一个不变的风险率。考虑关于死亡数量的列联表,其中一个维度是离散的时间,其他维度表示分类的解释变量。Holford(1980)以及 Laird 和 Olivier(1981)表明,关于该表的泊松分布对数线性模型及其似然函数,等价于假定生存时间具有分段指数风险(piecewise exponential hazards)的对数线性风险模型及其似然函数。

当时间区间很短时,分段指数法实质上是一种非参数方法,它不需要对风险随时间的变动做任何假定。这表明,可以通过由一个待指定的函数 $\lambda(t)$ 来替代 λ,对式 9.20 模型加以扩展,即

$$h(t;\mathbf{x}) = \lambda(t)\exp(\boldsymbol{\beta}'\mathbf{x})。$$

这就是 *Cox 比例风险模型*(*Cox proportional hazards model*)。在该模型中,风险之间的比率

$$h(t;\mathbf{x}_1)/h(t;\mathbf{x}_2) = \exp[\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)]$$

对所有的 t 都相同。

9.7.4　肺癌患者存活状况的例子*

表 9.13 描述了 539 名被诊断为肺癌的男性患者的存活情况。预测变量包括疾病的组织结构(histology)(H)和阶段(S)。在利用分段指数风险法时,将跟踪期(T)划分成了每两个月一段的区间。

令 μ_{ijk} 表示期望的死亡数,t_{ijk} 为在第 k 个区间内组织结构为 i 和疾病阶段 j 的病人的总历险时长。模型

$$\log(\mu_{ijk}/t_{ijk}) = \lambda + \lambda_i^H + \lambda_j^S + \lambda_k^T \tag{9.21}$$

具有残差 $G^2=43.9$($df=52$)。所有假定跟踪期区间与预测变量不存在交互效应的模型都是比例风险模型,因为在每个时间区间内组织结构和疾病阶段的效应都是相同的。表 9.14 给出了几个模型的拟合结果。尽管疾病阶段是一个很重要的预测因素,在控制了其他变量的情况下,组织结构的效应并不显著。

对于式 9.21 模型,疾病阶段的效应满足:

$$\hat{\lambda}_2^S - \hat{\lambda}_1^S = 0.470(SE = 0.174),$$

$$\hat{\lambda}_3^S - \hat{\lambda}_1^S = 1.324(SE = 0.152)。$$

例如,在固定的跟踪期内,对于给定的组织结构,疾病第三阶段所估计的死亡率是第一阶

段的 exp(1.324) =3.8 倍。在模型中加入疾病阶段和时间的交互项,并不能显著地改善拟合结果(G^2 的变动为 14.9,自由度的变动为 12)。在不包括组织结构效应的较简单模型中,$\{\hat{\lambda}_j^S\}$ 与上述结果非常相似。

表 9.13 死于肺癌的人数

跟踪期区间月	疾病阶段	疾病组织结构[a]								
		Ⅰ			Ⅱ			Ⅲ		
		1	2	3	1	2	3	1	2	3
0—2		9	12	42	5	4	28	1	1	19
		(157	134	212	77	71	130	21	22	101)
2—4		2	7	26	2	3	19	1	1	11
		(139	110	136	68	63	72	17	18	63)
4—6		9	5	12	3	5	10	1	3	7
		(126	96	90	63	58	42	14	14	43)
6—8		10	10	10	2	4	5	1	1	6
		(102	86	64	55	42	21	12	10	32)
8—10		1	4	5	2	2	0	0	0	3
		(88	66	47	50	35	14	10	8	21)
10—12		3	3	4	2	1	3	1	0	3
		(82	59	39	45	32	13	8	8	14)
12 +		1	4	1	2	4	2	0	2	3
		(76	51	29	42	28	7	6	6	10)

a 括号中的数字表示跟踪的总时长。

来源:由生物计量学会授权重印(the Biometric Society),基于:Holford(1980)。

表 9.14 对表 9.13 数据拟合具有比例风险的泊松回归模型的结果

效应[a]	G^2	df
T	170.7	56
$T+H$	143.1	54
$T+S$	45.8	54
$T+S+H$	43.9	52
$T+S+H+S\times H$	41.5	48

a T,跟踪期时间;H,疾病组织结构;S,疾病阶段。

9.7.5 对加权数据的分析*

拟合包括抵消项的对数线性模型的过程也可以应用于其他方面。对于期望频数 $\{\mu_i\}$ 和固定的常数 $\{t_i\}$,考虑以下模型:

$$\log(\mu_i/t_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots。$$

在标准的对数线性模型中,存在 $\{t_i=1\}$。这个一般形式的模型可用于分析抽样设计比简单随机抽样更复杂的分类数据。

许多调查的抽样设计都使用分层和/或整群抽样。按照抽样设计的特点,我们通过权数来扩大或缩小每个观测值的影响。对第 i 个单元格中的对象乘以权数就得到了关于该单元格的总的加权频数。平均单元格权数 z_i 被定义为单元格的加权频数除以它的未加权计数。以 $\{z_i\}$ 为条件,关于加权后的期望频数 $\{z_i\mu_i=\mu_i/t_i\}$ 的对数线性模型,其中 $t_i=z_i^{-1}$,可以表示为一个关于 $\{\log\mu_i\}$ 的标准对数线性模型,抵消项为 $\{\log t_i = -\log z_i\}$。通过拟合这个模型,可以得出适当的参数估计和标准误(Clogg and Eliason,1987)。

9.8 列联表模型分析中的空单元格和稀疏数据问题

在列联表中如果有些单元格的计数很小,我们称之为*稀疏*(*sparse*)数据。在本章的最后,我们讨论稀疏数据对模型拟合的影响。当样本规模 n 很小时,表格就会出现稀疏数据的问题。另外,如果 n 很大,但是单元格的数量也很大时,也可能会出现这一问题。对于包括许多个变量的表格,稀疏数据非常常见。下面的讨论适用于一般形式的列联表和模型,对于落在 N 个单元格中的 n 个观测值,单元格计数为 $\{n_i\}$,期望频数为 $\{\mu_i\}$。

9.8.1 空单元格:抽样性零值与结构性零值

稀疏表格通常会包含具有 $n_i=0$ 的单元格。这些*空单元格*(*empty cells*)可以分为两种:*抽样性零值*(*sampling zeros*)和*结构性零值*(*structural zeros*)。多数情况下,即便 $n_i=0$,仍然有 $\mu_i>0$,即该单元格有可能会存在观测值,如果 n 足够大的话,那么 $n_i>0$。这样的空单元格被称为*抽样性零值*。表 9.1 关于学生调查的数据中,空单元格就属于这种*抽样性零值*。

不可能存在观测值的空单元格被称为*结构性零值*。对于这样的单元格,$\mu_i=0$,而且无论 n 多大,必然存在 $\hat{\mu}_i=0$ 和 $n_i=0$。比如在一个按照性别、种族和癌症类别对癌症患者进行交叉划分的表格中,某些癌症(如前列腺癌,卵巢癌)是针对特定性别的。因而,一些空单元格是结构性零值。包括结构性零值的列联表被称为*不完整表格*(*incomplete tables*)。

抽样性零值是数据的一部分。对于一个泊松或多项分布变量来说,允许出现计数为 0 的情况。这些单元格也会被包括到似然函数和模型拟合的过程中。相反,结构性零值不是一个观测值,而且也不是数据的一部分。抽样性零值比结构性零值要普遍得多,因此以下的讨论主要针对前者的情况。

9.8.2 对数线性模型/Logit 模型估计值的存在性

在对数线性模型和 Logit 模型中,抽样性零值会影响模型参数的最大似然估计值的存在性。基于 Birch(1963)和 Fienberg(1970b)的早期工作,Haberman(1973b,1974a)对此进行了研究。令 $\mathbf{n}$ 表示单元格计数的向量,$\boldsymbol{\mu}$ 是它们的期望值。在泊松分布样本的情况下,Haberman 给出了结果 1 至 5,但是按照结果 6,它们也适用于多项分布样本。

1. 对数似然函数是一个关于 $\log\boldsymbol{\mu}$ 的严格凹函数。
2. 如果存在一个关于 $\boldsymbol{\mu}$ 的最大似然估计,那么它是唯一的并且满足似然方程 $\mathbf{X}'\mathbf{n}=\mathbf{X}'\hat{\boldsymbol{\mu}}$。相反,如果 $\hat{\boldsymbol{\mu}}$ 满足模型同时也满足似然方程,它就是关于 $\boldsymbol{\mu}$ 的最大似然估计。
3. 如果所有的 $n_i>0$,存在关于对数线性模型参数的最大似然估计。
4. 对于一个在某些边际表中观察计数与拟合计数相等的对数线性模型,假定存在对其参数的最大似然估计,那么这些边际表所有的计数都为正。
5. 如果关于模型 M 的最大似然估计存在,那么对于 M 的任何特例的最大似然估计也存在。
6. 对于任意对数线性模型,多项分布样本和独立的泊松分布样本下的最大似然估计 $\hat{\boldsymbol{\mu}}$ 是一致的,而且它们存在的条件也相同。

具体来说,考虑饱和模型的情况。按照结果 2 和结果 3,当所有的 $n_i>0$ 时,对 $\boldsymbol{\mu}$ 的最大似然估计为 $\mathbf{n}$。按照结果 4,当任意 $n_i=0$ 时,关于参数的最大似然估计不存在。模型的参数估计是针对 $\{\log \hat{\mu}_i\}$ 的,而且对于饱和模型有 $\hat{\boldsymbol{\mu}}=\mathbf{n}$,所以只有当所有的 $n_i>0$ 时,估计值才取有限值。

在非饱和模型的情况下,按照结果 3 和结果 4,当所有的 $n_i>0$ 时最大似然估计存在,而当与充分统计量有关的一组边际表中任意一个单元格计数为零时,最大似然估计不存在。假定至少有一个 $n_i=0$,但所有充分统计量的边际计数都为正,对于分级的对数线性模型,Glonek 等(1988)证明,当且仅当模型可分解时(注解 8.2),其中包括条件独立性模型,所有充分统计量的边际计数为正意味着最大似然估计存在。但是,对于包括所有成对关联项的模型,情况更复杂一些。以模型(XY,XZ,YZ)为例,当只有一个 $n_i=0$ 时,最大似然估计存在;当至少两个单元格为空时,最大似然估计可能就不存在。比如表9.15就不存在最大似然估计,尽管这里所有的充分统计量(二维的边际总计)都为正(习题 9.47)。

表 9.15　模型(XY,XZ,YZ)不存在最大似然估计的数据[a]

	Z:	1		2	
X	Y:	1	2	1	2
1		0	*	*	*
2		*	*	*	0

a 单元格中的 * 代表任意正数。

Haberman 表明,似然函数值的上确界(supremum)是有限值。这促使他界定了关于 $\boldsymbol{\mu}$ 的*扩展最大似然估计*(*extended ML*)。扩展最大似然估计在任何情况下都存在,但是它有可能会等于 0 或者落在边界上,而且它不一定具有普通最大似然估计值所具有的特性[另见:Baker et al.(1985)]。使模型成立且收敛于扩展估计的一系列估计值的对数似然值接近于其上确界。在这种扩展的意义上,$\hat{\mu}_i=0$ 是当 $n_i=0$ 时对饱和模型中的 μ_i 的最大似然估计,并且关于对数线性模型的参数估计可以是无穷大。

当某一变量的充分统计量的边际计数等于零时,对该项的估计值为无穷大。例如,当 XY 的一个边际总计等于零时,在对数线性模型(XY,XZ,YZ)中,$\{\hat{\lambda}_{ij}^{XY}\}$ 会出现无穷大;在分析 X 对 Y 的效应的 Logit 模型中,$\{\hat{\beta}_i^X\}$ 也会出现无穷大估计。甚至在某些情况下,无穷大估计也不存在。这样的一个例子是在 2×2 表格中行或列的两个计数都为零时,估计其对数发生比之比。

对参数的最大似然估计值为∞(或 $-\infty$)意味着,在一些单元格中最大似然拟合值等于 0,并且对某些发生比之比的估计值等于∞ 或 0。当迭代拟合过程不收敛时,就表明可能存在这种情况,这往往是因为估计值从一个循环到下一个循环不断上升。但是,在迭代过程达到某个点后,很多软件会被接近水平的似然函数所欺骗,报告结果已经收敛。此时,由于对数似然函数的曲率非常小,它所估计的标准误(基于由二阶偏导数组成的信息矩阵的逆矩阵)特别大并且极不稳定。这时,数据发生微小的变化常常引起估计值及其标准误的明显变动。在稀疏数据下的一个问题就是,大家可能没有意识到真实的拟合值是无穷大,结果给出了无效的而且极不稳定的估计值以及相应的统计推断。

许多最大似然分析不受空单元格的影响。即使当某一参数的估计值为无穷大时,它对数据分析而言并不是致命的。此时,关于真实的发生比之比的似然比置信区间以无穷大作为它的一个端点。例如,在一个二维表格中,当 $n_{11}=0$ 但其他 $n_{ij}>0$ 时,$\log\hat{\theta}=$

$-\infty$,它的置信区间为$(-\infty, U)$,其中 U 表示一个取有限值的上界。然而,当空单元格的分布模式导致模型的某些拟合值必定为 0 时,这会影响进行模型拟合检验的自由度(Haslett, 1990)。

9.8.3 例子:临床试验的数据

表 9.16 所示为一项在五个中心进行的临床试验的结果。研究目的是比较一种现行药与安慰剂治疗真菌感染的效果,结果变量为二分变量(成功,失败)。对于该数据,令 Y = 结果变量,X = 干预方式($x_1 = 1$ 表示现行药,$x_2 = 0$ 表示安慰剂),Z = 中心。

表 9.16 临床试验的干预方式与结果以及 XY 和 XZ 边际表[a]

中心	干预方式	结果		YZ 边际	
		成功	失败	成功	失败
1	现行药	0	5	0	14
	安慰剂	0	9		
2	现行药	1	12	1	22
	安慰剂	0	10		
3	现行药	0	7	0	12
	安慰剂	0	5		
4	现行药	6	3	8	9
	安慰剂	2	6		
5	现行药	5	9	7	21
	安慰剂	2	12		
XY	现行药	12	36		
边际	安慰剂	4	42		

a X,干预方式;Y,结果;Z,中心。

来源:数据由山德士制药公司(Sandoz Pharmaceuticals Corporation)的 Diane Connell 友情提供。

第 1 个和第 3 个中心没有出现成功的案例。因此,在对干预方式进行合并后的关于结果变量与中心的 5 × 2 边际表格中包含空单元格。表 9.16 的最后两列给出了这个边际表。对此数据,在包括 YZ 关联项的对数线性模型或 Logit 模型中会出现参数的最大似然估计为无穷大的情况。以如下的 Logit 模型为例:

$$\text{Logit}[P(Y=1 \mid X=i, Z=k)] = \beta x_i + \beta_k^Z。$$

(在此我们忽略了截距项,所以不需要对$\{\beta_k^Z\}$加以限定;这时,相应参数表示的是中心的效应,而不是各中心与一个作为基线的中心之间的对比)。随着 β_1^Z 和 β_3^Z 向 $-\infty$ 递减,似然函数持续上升;也即,当 Logit 下降到 $-\infty$ 时,模型对这些中心所拟合的成功概率下降到最大似然估计值 0。

如表 9.16 的底部所示,结果变量和干预方式之间的 2 × 2 边际表中的计数都为正值。在这个 Logit 模型中,表 9.16 的空单元格会影响对中心效应的估计,但不会影响对干预效应的估计。当对数似然函数上升到极限值时,模型所拟合的对数发生比之比为 $\hat{\beta} = 1.55$ (SE = 0.70)。多数软件会报告这一结果,但是它们不会报告 $\hat{\beta}_1^Z = \hat{\beta}_3^Z = -\infty$ 而是给出很大的估计值和极大的标准误。例如,SAS 的 PROC GENMOD 给出的关于 $\hat{\beta}_1^Z$ 和 $\hat{\beta}_3^Z$ 的值大约为 −26,相应的标准误约为 200 000。

在上述数据中删除第 1 个和第 3 个中心,也可以得到关于干预效应的估计 $\hat{\beta} = 1.55$。

当一个中心的所有观测值都落在结果变量的某一个类别时，它对估计这个发生比之比没有意义（但它确实会影响一些其他的关联指标，如比例之差）。事实上，这样的分表对标准的条件独立性检验也没有帮助，比如 Cochran-Mantel-Haenszel 检验（第 6.3.2 节）和精确检验（第 6.7.5 节）。

在多中心分析中，另一种策略是将类型相似的中心加以合并。这时，如果在每个合并后的分表中结果变量的两种情况都有发生，统计推断就会使用所有的数据。对于表 9.16，可能第 1 个和第 3 个中心与第 2 个中心相似，因为后者的成功率也非常低。将这些中心进行合并后，再重新拟合模型得到 $\hat{\beta}=1.56(SE=0.70)$。通常来说，这种方法所得到的结果与直接删除掉那些结果变量只出现在一个类别中的分表很相似。

9.8.4 小样本对 X^2 和 G^2 的影响

尽管空单元格和稀疏表格不一定会影响所关注的参数估计，它们会导致拟合优度统计量的样本分布远远不同于卡方分布。对于单元格数量 N 固定的表格，随着 $n\rightarrow\infty$，这些统计量的真实样本分布收敛于卡方分布。统计量对卡方分布近似的充分性既取决于 n，也取决于 N。

Cochran 在几篇文章中讨论了关于 X^2 的卡方分布近似。在 1954 年，他建议在 $df>1$ 的独立性检验中，只要不超过 20% 的 $\mu_i<5$，最小的期望值 $\mu_i\approx1$ 也是可以接受的。Koehler（1986）、Koehler 和 Larntz（1980）以及 Larntz（1978）表明，对于较小的 n 和稀疏表格，X^2 的适用性比 G^2 更强。当 n/N 小于 5 时，G^2 分布对卡方的近似通常很差。根据稀疏数据的具体情况，基于 G^2 服从卡方分布假设所得到的 P 值可能过大也可能过小。当大多数的 μ_i 小于 0.5 时，将 G^2 视为服从卡方分布所得到的检验会非常保守：当 H_0 为真时，G^2 检验报告的 P 值一般远远大于其真实值。当大多数的 μ_i 在 0.5 到 4 之间时，G^2 检验又太宽松：所报告的 P 值往往小于其真实值。

随着 N 的增加，能使 X^2 对卡方分布充分近似的 n/N 的大小一般会下降（Koehler and Larntz，1980）。然而，对于既包含很小的 μ_i 又包含较大的 μ_i 的稀疏表格，这种近似结果比较差（Haberman，1988）。我们难以给出一个涵盖各种情况的一般性指导原则。有关对此问题的其他讨论，参见：Cressie and Read（1989）、Lawal（1984）。

对于固定的 n 和 N，自由度较小的检验对卡方的近似会好一些。例如，在 $I\times J\times K$ 表格中进行条件独立性检验时，$G^2[(XZ,YZ)\mid(XY,XZ,YZ)]$（自由度为 $df=(I-1)(J-1)$）比 $G^2[(XZ,YZ)]$（自由度为 $df=K(I-1)(J-1)$）更近似于卡方分布。在同质的双线性 XY 关联模型中，对 $H_0:\beta=0$ 的定序检验的自由度为 $df=1$，它对卡方的近似更好。

9.8.5 基于模型的检验与稀疏数据

根据式 9.3 和式 9.4，基于模型的统计量 $G^2(M_0\mid M_1)$ 和 $X^2(M_0\mid M_1)$ 对数据的依赖只取决于拟合值，因而它们只与较复杂模型的最小充分统计量有关。随着最小充分统计量的期望值上升，这些统计量的零分布收敛于卡方分布。对于大多数对数线性模型而言，这些充分统计量与边际表有关，而边际总计比单个的单元格计数更接近于正态分布。因此，$G^2(M_0\mid M_1)$ 和 $X^2(M_0\mid M_1)$ 收敛于极限卡方分布的速度要比 $G^2(M_0)$ 和 $X^2(M_0)$ 更快，后两个统计量依赖于单个的单元格计数。

当 $\{\hat{\mu}_i\}$ 很小但是有关 M_1 的充分统计量中边际总计大多至少在 5 到 10 之间时，进行模型比较的统计量对卡方的近似一般都是充分的。Haberman（1977a）在理论上对此进行了论证。

9.8.6 其他渐近性和统计量

当大样本近似不充分时,可以考虑精确小样本方法。如果精确方法本身不可行,常常有可能通过蒙特卡洛法对精确分布进行很好的近似(如:Booth and Butler (1999); Forster et al. (1996); Kim and Agresti (1997); Mehta et al. (1988))。

另一种方法是稀疏渐近近似法,它适用于单元格数量 N 随 n 的增加而增加的情况。与通常的大样本理论(固定的 $N, n\to\infty$)不同,在这种方法中,$\{\mu_i\}$ 不一定会增加。Koehler 和 Larntz(1980)表明,在对一个指定的多项分布模型进行拟合优度检验时,对于非常稀疏的表格,存在一个标准化形式的 G^2 近似服从正态分布。McCullagh(1986)综述了处理稀疏表格的有关方法,并介绍了另一种关于 G^2 的近似方法。Zelterman(1987)给出了关于 X^2 的正态近似,并提出了相应的统计量。

9.8.7 在列联表单元格中加入常数

空单元格和稀疏表格会导致各种问题,如对数线性模型的参数估计可能不存在、无法估计发生比之比、运算算法不能正常进行以及卡方统计量的渐近近似不充分。但是,这些问题并不一定就会影响数据分析。比如,仍然可以对似然函数进行最大化,点估计为 ∞ 的效应似然比置信区间通常包括一个有限值的下界,并且可以使用小样本推断方法来取代渐近方法。

一种保证所有的效应估计都是有限值并且拟合算法收敛的办法是,对每个单元格计数加上一个很小的常数,如同 Goodman(1964b,1970,1971a)在拟合饱和模型时所建议的那样,一些算法在每个单元格中加入$\frac{1}{2}$。对于饱和模型,这样处理是有好处的,例如在估计 2×2 表格的发生比之比时它能消减偏差(Gart,1966;Gart and Zweiful,1967)。然而,在拟合非饱和模型之前,对每个单元格加上$\frac{1}{2}$会造成对数据的过分修匀,从而破坏了样本的分布。这种方法会导致效应估计和检验统计量过于保守。当单元格数量很大时,这种影响是非常严重的。

即使对于饱和模型,在每个单元格中加上$\frac{1}{2}$也不是解决所有问题的万灵丹。当某一发生比之比的普通最大似然估计为无穷大时,对每个单元格加入$\frac{1}{2}$后会确保估计值为有限值,同时置信区间的所有端点也是有限值。但是,这种情况下使用 ∞ 作为发生比之比的上界更合理,因为样本中没有任何证据表明发生比之比会小于某一给定的值。

当我们对稀疏数据的影响存在疑虑时,应当考虑进行灵敏度分析(sensitivity analysis)。例如,对于每个可能影响结果的观测值,看看删除它或把它移到另一个单元格等对数据的微调会不会导致分析结果发生变化。在这种情况下,关于广义线性模型的影响力诊断(Williams,1987)也很有帮助。通常来说,一些关联不受空单元格的影响并且在各种分析中的结果也很稳定,而另外一些关联可能会受到影响并且结果极不稳定。如果对数据的微调会对关联产生影响,那么在下结论时应当非常谨慎。

在后面的章节中,我们将介绍一些比在单元格中加入随意常数更一般化的数据修匀方法,其中包括随机效应模型(第 12.3 节)和贝叶斯方法(第 15.2 节)。

注 解

第9.1节:关联图与可合并性

9.1 Darroch 等(1980)界定了一组*图示模型*(*graphical models*),其中包括可分解模型(参见注解8.2)。有关对图示模型以及相应的*独立图*(*independence graphs*)——展示条件独立性的结构——的介绍,另见:Anderson and Böckenholt(2000)、Edwards(2000)、Edwards and Kreiner(1983)、Kreiner(1998)、Lauritzen(1996)、Whittaker(1990)。Whittaker(1990, sec. 12.5)总结了各种可合并性定义之间的关系。

9.2 对于 $I\times J\times 2$ 表格,可合并性条件(第9.1.2节)既是必要的,也是充分的(Simpson, 1951; Whittemore, 1978)。对于 $I\times J\times K$ 表格,Ducharme 和 Lepage(1986)证明,这些条件是无论怎样合并 Z 的类别(即,对 Z 进行各种各样的部分合并),发生比之比都不变的充要条件。

Darroch(1962)定义了一个*完美*(*perfect*)表格,其中对于任何的 i,j,k,

$$\sum_i \frac{\pi_{ij+}\pi_{i+k}}{\pi_{i++}} = \pi_{+j+}\pi_{++k}, \qquad \sum_j \frac{\pi_{+jk}\pi_{ij+}}{\pi_{+j+}} = \pi_{i++}\pi_{++k},$$

$$\sum_k \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} = \pi_{i++}\pi_{+j+}。$$

在完美表格中,同质性关联意味着

$$\{\pi_{ijk} = \pi_{ij+}\pi_{i+k}\pi_{+jk}/\pi_{i++}\pi_{+j+}\pi_{++k}\},$$

并且条件发生比之比等于边际发生比之比。Whittemore(1978)利用完美表格显示,对于 $K>2$ 的 $I\times J\times K$ 表格,即使任意一对变量都不条件独立,条件发生比之比和边际发生比之比仍有可能相等。另见:Davis(1986b)。

假定在 Z 的每个取值水平上所计算的关于二分结果变量 Y 与预测变量 X 之间的比例之差或相对风险都相等,如果在 XZ 边际表中 Z 独立于 X 或者如果在给定 X 后 Z 条件独立于 Y,那么该指标在 XY 边际表中取值相同(Shapiro, 1982)。因此,对于在每个取值组合处都存在相同数量的观测值的析因设计(factorial design),比例之差和相对风险是可合并的。另见:Wermuth(1987)。

第9.2节:模型选择与比较

9.3 有关对数线性模型选择的文章包括:Aitkin(1979, 1980)、Benedetti and Brown(1978)、Brown(1976)、Goodman(1970, 1971a)、Wermuth(1976)、Whittaker and Aitkin(1978)。当某个模型成立时,G^2/df 的渐近均值等于1。Goodman(1971a)建议在比较模型拟合情况时使用这一指标。它的值越小表示拟合得越好。

9.4 Kullback 等(1962)以及 Lancaster(1951)是关于分割多维表的卡方统计量的最早研究者之一。Goodman(1970)以及 Plackett(1962)指出了他们的方法所存在的问题。当观测值服从自然指数族分布时,Simon(1973)证明,只要模型是自然参数的线性形式,就有 $G^2(M_0 \mid M_1) = 2\sum_i \hat{\mu}_{1i}\log(\hat{\mu}_{1i}/\hat{\mu}_{0i})$。有关更复杂模型的分割,参见:Lang(1996b)。

第9.4节:对定序关联的模型分析

9.5 Goodman(1979a)激发了有关定序数据的对数线性模型的研究。他扩展了 Haberman(1974b)的早期工作,即将 λ^{XY} 关联项表示为正交多项式的形式。对于多维表中的更一般性的定序模型,参见:Agresti(1984)、Becker(1989a)、Becker and Clogg(1989)、Goodman(1986)。

第 9.6 节：关联模型、相关模型与对应分析

9.6 关于 *RC* 模型的早期研究包括：Goodman(1979a,1981a,b)、Andersen(1980,pp. 210-216)，他们明显部分受到了 G. Rasch 的早期工作的启发(参见：Anderson (1995))。对在多维表情况下的扩展，参见：Anderson and Böckenholt(2000)、Becker(1989a,b,1990)、Becker and Clogg(1989)、Chuang et al. (1985)、Goodman(1985,1986,1996)。Anderson(1984)介绍了一个与此有关的模型。Anderson 和 Vermunt(2000)表明，当观察变量在给定一个服从条件正态分布的潜变量后条件独立时，就会推导出 *RC* 模型以及相应的关联模型。他们的研究是对 Lauritzen 和 Wermuth (1989) 以及 Whittaker 关于 van der Heijden 等(1989)的有关评论的扩展。另见：de Falguerolles et al. (1995)。Clogg 和 Shihadeh(1994)对关联模型以及相应的相关模型进行了综述。

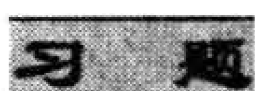
9.7 Kendall 和 Stuart(1979，第 33 章)综述了关于列联表的主要典型相关模型。另见 Williams(1952)，他讨论了 R. A. Fisher 和其他人的早期工作。Karl Pearson 常利用一个潜在的二元正态分布假定来对表格进行分析(第 16.1 节)。有关对该分布的相关系数的估计，参见：Becker(1989b)、Goodman(1981b)、Kendall and Stuart(1979, chaps. 26 and 33)、Lancaster(1969,chap. 10)、Pearson(1904)关于 2×2 表格的四项相关系数、Lancaster and Hamdan(1964)(关于 $I\times J$ 表格的多项相关系数)。

9.8 在 Benzécri(如参见(Benzécri,1973))的影响下，对应分析在法国得到了广泛应用。Goodman(1996)将对应分析归功于 H. O. Hartley，文章使用其原来的德文名字发表(Hirschfeld, 1935)。Greenacre(1993)将其与矩阵的奇异值分解联系起来。其他的有关讨论，参见：Escoufier(1982)、Friendly(2000, chap. 5)、Goodman(1986,1996,2000)、Michailidis and de Leeuw(1998)、van der Heijden and de Leeuw(1985)、van der Heijden et al. (1989)。Gabriel(Gabriel,1971)讨论了有关 biplots 的相应研究。

第 9.7 节：关于比率的泊松回归

9.9 关于抵消项的另一项应用是对表格的标准化(第 8.7.4 节)。有关比率数据的分析，参见：Breslow and Day (1987, sec 4. 5)、Freeman and Holford (1980)、Frome (1983)、Hoem(1987)。有关讨论分组的生存数据的文章，尤其是关于存活概率的对数线性模型和 Logit 模型，包括：Aranda-Ordaz(1983)、Larson(1984)、Prentice and Gloeckler(1978)、Schluchter and Jackson (1989)、Stokes et al. (2000, chap. 17)、Thompson(1977)。Aitkin 和 Clayton(1980)讨论了指数生存模型，并介绍了风险函数服从威布尔或极值生存分布的相应模型。对数似然函数式 9.19 实际上只适用于无信息(*noninformative*)的删失情况。当研究对象基于与研究本身有关的考虑从该研究中退出时，如担心某种干预方式可能会影响健康，这种假定是不合理的。

9.10 Lindsey 和 Mersch(1992)介绍了一种利用对数线性模型拟合形式为式 4.14 的指数族分布 $f(y;\theta)$ 的聪明方法，其中 ϕ 是已知的。该方法将结果变量分割成区间 $\{(y_k-\Delta_k/2, y_k+\Delta_k/2)\}$。落在这些区间中的计数服从概率近似于 $\{f(y_k,\theta)\Delta_k\}$ 的多项分布。在存在抵消项的情况下，期望频数的对数近似等于有关 θ 的一个线性函数。



习题

应用部分

9.1 利用表 8.3 中的发生比之比展示可合并性条件。

a. 对于(A,C,M),所有的条件发生比之比都等于1.0。说明为什么所给出的边际发生比之比也都等于1.0。

b. 对于(AC,M),说明为什么(i)所有的条件发生比之比都等于边际发生比之比,(ii)所有的$\hat{\mu}_{ac+}=n_{ac+}$。

c. 对于(AM,CM),说明为什么(i)值为1.0的AC条件发生比之比不一定与其边际发生比之比相等,(ii)AM和CM的条件发生比之比等于它们的边际发生比之比,(iii)所有的$\hat{\mu}_{a+m}=n_{a+m}$和$\hat{\mu}_{+cm}=n_{+cm}$。

d. 对于(AC,AM,CM),说明为什么(i)所有的条件发生比之比都不一定等于相应的边际发生比之比,(ii)边际发生比之比的拟合值一定等于样本的边际发生比之比。

9.2 表9.17是对一项研究的总结,该研究包括的变量为母亲的年龄(A)、以天数计算的妊娠期长度(G)、婴儿存活状况(I),以及母亲在孩子出生前每天的吸烟数(S)。将G和I作为结果变量,A和S作为解释变量。

表9.17 习题9.2的数据

年龄	吸烟量	妊娠期长度	婴儿存活状况	
			否	是
<30	<5	≤260	50	315
		>260	24	4 012
	5 +	≤260	9	40
		>260	6	459
30 +	<5	≤260	41	147
		>260	14	1 594
	5 +	≤260	4	11
		>260	1	124

来源:N. Wermuth, pp. 279-295 in *Proc. 9 th International Biometrics Conference*, vol. 1(1976)。经生物计量学会(the Biometric Society)授权重印。

a. 说明为什么对数线性模型应当包括λ^{AS}项。

b. 拟合模型$(AGIS)$、(AGI,AIS,AGS,GIS)、(AG,AI,AS,GI,GS,IS),以及(AS,G,I)。指出嵌套于上述任两个模型之间的一组对数据拟合很好的模型,从中选取一个模型。

c. 通过(i)前向选择法,(ii)后向剔除法来构建一个模型。比较这两种不同方法所给出的结果,并对所选定的模型加以解释。

9.3 参见表2.13。考虑一组嵌套模型$\{(DVP),(DP,VP,DV),(VP,DV),(P,DV),(D,V,P)\}$。通过卡方分割来分别比较四对模型,保证在四次比较中总的第一类错误的概率不超过$\alpha=0.10$。由模型(DVP)开始,通过后向选择法,你会选择哪个模型?对模型(DP,VP,DV)、(DP,DV)、(P,DV)和(D,V,P)重复上述分析,表明最终所选中的模型取决于嵌套模型的可选集合。

9.4 对表6.3的数据选择对数线性模型。

a. 为什么忽略λ^{GM}项的对数线性模型是不合理的?

b. 由模型(GM,E,P)开始,利用前向选择法,表明模型(GM,GP,EG,EMP)是合理的。

c. 利用后向剔除法,表明(*GM*,*GP*,*EMP*)或(*GM*,*GP*,*EG*,*EMP*)是合理的。

d. *EMP* 交互项看起来很关键。为了对此加以描述,表明对于不存在婚前性行为的对象,婚外性行为对离婚的效应更大。

e. 利用残差来描述模型(*GM*,*EMP*)拟合不充分的情况。

9.5 在关于表 8.3 的模型(*AC*,*AM*,*CM*)中,每个单元格的标准化皮尔逊残差等于 ±0.63。解释为什么每个残差的绝对值都一样。相反,模型(*AM*,*CM*)在当 *M* = "是"(如当 *A* = *C* = "是"时为 +3.70)时每个单元格的标准化皮尔逊残差为 ±3.70,当 *M* = "否"(如当 *A* = *C* = "是"时为 +12.80)时每个单元格残差为 ±12.80。加以解释。

9.6 参考表 8.8。对一个不包括三维交互项的模型进行残差分析,并描述交互效应的特点。

9.7 对表 3.2 数据对应的独立性模型进行残差分析。说明为什么结果表明,双线性关联模型可能拟合得更好。拟合此模型,将其与独立性模型进行比较,并加以解释。

9.8 参考习题 9.7。

a. 利用标准化的赋值,求 $\hat{\beta}$。对关联的强度进行评述。

b. 拟合一个以工作满意度的赋值为参数的模型。解释所估计的赋值,并将拟合结果与 $L \times L$ 模型加以比较。

9.9 参考表 9.3。

a. 对于双线性关联模型,构建由四个角部的单元格所形成的发生比之比的 95% 的置信区间。加以解释。

b. 拟合列效应模型。将所估计的列赋值与(a)部分所使用的等距赋值加以对比。给定模型成立时,检验真实的列赋值是否等距,并加以解释。构建关于四个角部单元格的发生比之比的 95% 的置信区间。与(a)部分的结果进行比较。

9.10 对于不相邻的类别而言,一个很弱的局部关联可能是非常重要的。通过关于表 9.9 的 $L \times L$ 模型加以展示,指出与所估计的局部发生比之比相比,关于四个角部单元格的发生比之比的估计值是多少。

9.11 参考表 7.8。拟合一个同质的双线性关联模型,并加以解释。在控制性别(*G*)后,在以下模型中分别检验收入(*I*)与工作满意度(*S*)之间的条件独立性:(a)上述模型,(b)模型(*IS*,*IG*,*SG*)。说明为什么结果会如此不同。

9.12 对表 9.3 中的数据拟合 *RC* 模型。解释所估计的赋值。这个模型比一致性关联模型拟合得更好吗?

9.13 对于表 9.9 的数据,重做第 9.6 节中的相关模型和对应分析的结果。

9.14 一百名白血病患者被随机分配到两个干预组。在研究期间,干预组 A 中有 10 人死亡,干预组 B 中有 18 人死亡。干预组 A 的总历险时长为 170.4 年,而干预组 B 的总历险时长为 147.3 年。检验两个干预组的死亡率是否相同。利用置信区间对两个死亡率加以比较。

9.15 对于表 9.11,拟合一个死亡率只依赖于年龄的模型。对年龄效应加以解释。

9.16 考虑式 9.18 模型。以下变化如何影响模型的参数估计、它们的标准误以及拟合优度统计量:(a)历险时间翻番但死亡发生数不变;(b)历险时间不变但死亡发生数翻番;(c)历险时间和死亡发生数都翻番。

9.17 考虑表 9.13。说明如何分析风险函数是否随时间的变化而变化。

9.18 一篇由 W. A. Ray 等(*Amer. J. Epidemiol.* 132:873-884, 1992)所写的文章讨论年龄在 65~84 岁之间的 16 262 名研究对象的机动车事故发生率,该研究对每个对象进行了长达 4 年的跟踪调查。在女性研究对象的 17 300 年的总观察期中,发生了 175 次导致受伤的事故。在男性研究对象的 21 400 年的总观察期中,发生了 320 次导致受伤的事故。

a. 给出关于总的受伤事故发生率的 95% 的置信区间。

b. 通过模型比较男性和女性的事故发生率。

9.19 本书网站(*www. stat. ufl. edu/ ~ aa/cda/cda. html*)上的一个表格显示了 1970 至 1984 年间英国客运列车的总里程数(以百万英里为单位)以及相撞事故的数量。假定在 14 年间事故发生率的对数为常数 α 的一个泊松模型的拟合结果为 $\hat{\alpha} = -4.177$(SE = 0.132 5),$X^2 = 14.8$(df = 13)。加以解释。

9.20 表 9.18 列出了英国足球联盟二级联赛的球队在 1987—1988 赛季的观众总数(以千人计)与被捕总人数。令 Y = 一支球队的被捕数,令 t = 观众总数。说明为什么模型 $E(Y) = \mu t$ 可能成立。假定这是一个泊松分布样本,拟合该模型并解释结果。将被捕数与观众数进行绘图,并在图中标出预测方程。利用残差指出被捕数的观察值远远不同于期望值的球队。

表 9.18 习题 9.20 的数据

球队	观众数千人	被捕数	球队	观众数千人	被捕数
Aston Villa	404	308	Shrewsbury	108	68
Bradford City	286	197	Swindon Town	210	67
Leeds United	443	184	Sheffield Utd.	224	60
Bournemouth	169	149	Stoke City	211	57
West Brom	222	132	Barnsley	168	55
Hudderfield	150	126	Millwall	185	44
Middlesbro	321	110	Hull City	158	38
Birmingham	189	101	Manchester City	429	35
Ipswich Town	258	99	Piymouth	226	29
Leicester City	223	81	Reading	150	20
Blackburn	211	79	Oldham	148	19
Crystal Palace	215	78			

来源:《独立报》(伦敦),1988 年 12 月 21 日。感谢 P. M. E. Altham 向我提供了以上数据。

9.21 表 9.19 取自一项关于英国医生的研究。

表 9.19 习题 9.21 的数据

年龄	人年数		冠心病死亡数	
	非吸烟者	吸烟者	非吸烟者	吸烟者
35~44	18 793	52 407	2	32
45~54	10 673	43 248	12	104
55~64	5 710	28 612	28	206
65~74	2 585	12 663	28	186
75~84	1 462	5 317	31	102

来源:R. Doll and A. B. Hill, *Natl. Cancer Inst. Monogr.* 19:205-268(1996)。另见:N. R. Breslow in *A Celebration of Statisics*, ed. A. C. Atkinson and S. E. Fienberg, (New York:Springer-Verlag, 1985)。

a. 对每个年龄组,给出样本中非吸烟者和吸烟者每 1 000 人年数的冠心病死亡率。为了对其进行比较,求它们之间的比率,并描述该比率与年龄的关系。

b. 拟合一个关于该死亡率的对数的主效应模型，包括四个关于年龄的参数和一个关于吸烟的参数。在分析模型拟合不足的过程中，指出这个模型假定非吸烟者和吸烟者的冠心病死亡率之比不随年龄的变动而变动。

c. 根据(a)部分，说明为什么加入关于年龄和吸烟之间的交互项(将年龄作为连续变量)是合理的。对于该模型，表明冠心病死亡率之比的对数与年龄呈线性关系。对年龄组进行赋值，拟合该模型并解释结果。

9.22　利用定序 Logit 模型分析表 9.9。解释结果，并讨论这个模型与对数线性模型相比有哪些优缺点。

9.23　参考习题 8.6。利用本章所介绍的方法分析这些数据。

理论与方法

9.24　在一个 $2\times2\times K$ 表格中，XY 的条件发生比之比都相等，但却不等于 XY 的边际发生比之比。这里存在三维交互效应吗？Z 是否条件独立于 X 或 Y？加以说明。

9.25　考虑对数线性模型(WX, XY, YZ)。说明为什么在仅给定 X、仅给定 Y 或给定 X 和 Y 后，W 和 Z 之间独立。在什么情况下 W 和 Y 条件独立？在什么情况下 X 和 Z 条件独立？

9.26　假定对数线性模型(XY, XZ)成立。

a. 求 μ_{ij+} 和 $\log\mu_{ij+}$。证明关于 XY 边际表的对数线性模型具有与模型(XY, XZ)中的$\{\lambda_{ij}^{XY}\}$相同的关联参数。推导 XY 边际表的发生比之比与分表中的相同。利用关于模型(XY, XZ)的一个相似结果，推导第 9.1.2 节的可合并性条件。

b. 计算模型(XY, XZ, YZ)中的 $\log\mu_{ij+}$，并说明为什么边际关联并不一定等于条件关联。

9.27　对于一个四维表格，在以下对数线性模型中：(a)(WX, XYZ)，(b)(WX, WZ, XY, YZ)，WX 的条件关联等于它们的边际关联吗？为什么？

9.28　对数线性模型 M_0 是 M_1 的一个特例。

a. 说明为什么对于模型 M_0 的充分统计量的边际分布，两个模型的拟合值相等。

b. Haberman(1974a)证明，当$\{\hat{\mu}_i\}$满足模型 M_0 的任意一个特例时，存在 $\sum_i \hat{\mu}_{1i}\log\hat{\mu}_i = \sum_i \hat{\mu}_{0i}\log\hat{\mu}_i$。因此，$\hat{\boldsymbol{\mu}}_0$ 是 $\hat{\boldsymbol{\mu}}_1$ 对满足 M_0 的$\{\log\boldsymbol{\mu}\}$的线性组合的正交映射。利用该结果，证明 $G^2(M_0) - G^2(M_1) = 2\sum_i \hat{\mu}_{1i}\log(\hat{\mu}_{1i}/\hat{\mu}_{0i})$。

9.29　参考第 9.2.4 节。证明 $G^2(M_j \mid M_{j-1})$ 等于对由合并第 1 到第 $j-1$ 列与第 j 列所形成的 2×2 表格进行独立性检验的 G^2。

9.30　对于 T 个分类变量 $X_1, \cdots, X_T$，说明为什么：

a.

$$G^2(X_1, X_2, \cdots, X_T) = G^2(X_1, X_2) + G^2(X_1, X_2, X_3) + \cdots + G^2(X_1, X_2, \cdots, X_{T-1}, X_T)。$$

b.

$$G^2(X_1, \cdots X_{T-1}, X_T) = G^2(X_1, X_T) + G^2(X_1X_T, X_1X_2) + \cdots + G^2(X_1X_2\cdots X_{T-1}, X_1X_2, \cdots, X_{T-2}X_T)。$$

9.31　对于 $I\times2$ 的列联表，说明为什么双线性关联模型等价于线性 Logit 模型式 5.5。

9.32　考虑将$\{v_j = j\}$替代为$\{v_j = 2j\}$的式 9.6$L\times L$ 模型。说明为什么 $\hat{\beta}$ 等于原来的一半，但是$\{\hat{\mu}_{ij}\}$、$\{\hat{\theta}_{ij}\}$以及 G^2 都不变。

9.33 如果在任意 $x_1<x_2$ 且 $y_1<y_2$ 时，XY 的联合密度函数都满足 $f(x_1,y_1)f(x_2,y_2)\geqslant f(x_1,y_2)f(x_2,y_1)$，Lehmann（1966）将（$X,Y$）定义为正向似然比相依（*positively likelihood-ratio dependent*），那么，$Y(X)$ 的条件分布随着 $X(Y)$ 的上升而随机上升（Goodman，1981a）。

a. 对于 $L\times L$ 模型，证明 Y 和 X 的条件分布是随机排序的。在 $\beta>0$ 时，它有什么特点？

b. 在式 9.8 行效应模型中，如果 $\mu_i>\mu_h$，证明在第 i 行中 Y 的条件分布随机高于第 h 行。说明为什么 $\mu_1=\cdots=\mu_I$ 等价于各行中的 I 个条件分布相等。

9.34 如果一个表格中行和列的排序确保所有的局部对数发生比之比都非负，Yule（1906）将这样的表格定义为等方的（*isotropic*）（另见：Goodman（1981a））。

a. 证明如果一个表格满足以下模型，那么它就是等方的：(i) 双线性关联模型，(ii) 行效应模型，(iii) RC 模型。

b. 说明为什么将一个等方表格的相邻行或列进行合并后，它仍然是等方的。

9.35 考虑双线性关联模型的对数似然函数。

a. 将其对 β 求导，并给出在 $\beta=0$ 时的值以及关于参数的零分布估计，证明赋值函数与

$$\sum_i\sum_j u_iv_j(p_{ij}-p_{i+}p_{+j})$$

成比例。

b. 利用 δ 方法证明，它的零分布标准误等于

$$\{[\sum u_i^2p_{i+}-(\sum u_ip_{i+})^2][\sum v_j^2p_{+j}-(\sum v_jp_{+j})^2]/n\}^{1/2}。$$

c. 构建关于独立性检验的计分统计量。证明它本质上相当于相关系数检验式 3.15（Hirotsu（1982）讨论了关于定序备择假设的一组计分检验）。

9.36 根据习题 7.33 中的附加结果，证明如果式 7.24 累积 Logit 模型成立并且 $|\beta|$ 很小，那么具有行赋值 $\{x_i\}$ 和"ridit"列赋值 $\{v_j=[P(Y\leqslant j-1)+P(Y\leqslant j)]/2\}$ 的双线性关联模型应当会拟合得很好，并且它的 β 参数大约等于式 7.24 模型中 β 的两倍。

9.37 考虑式 9.8 行效应模型。

a. 证明令 $\lambda_I^X=\lambda_J^Y=\mu_I=0$ 不会影响模型的一般性。

b. 证明该模型的最小充分统计量是 $\{n_{i+}\}$、$\{n_{+j}\}$ 以及 $\{\sum_j v_jn_{ij},i=1,\cdots,I\}$，并推导它的似然方程。

9.38 证明列效应模型对应着一个 Y 是 X 的赋值的线性函数的基线类别 Logit 模型，其斜率取决于结果变量的成对类别。

9.39 参考同质的双线性关联模型式 9.10。

a. 证明其似然方程为，对于所有的 i,j 和 k，

$$\hat{\mu}_{i+k}=n_{i+k},\quad\hat{\mu}_{+jk}=n_{+jk},\quad\sum_i\sum_j u_iv_j\hat{\mu}_{ij+}=\sum_i\sum_j u_iv_jn_{ij+}。$$

b. 证明它的残差自由度为 $df=K(I-1)(J-1)-1$。

c. 当 $I=J=2$ 时，说明为什么它等价于（XY,XZ,YZ）。

d. 说明在异质的双线性 XY 关联式 9.11 情况下，上面最后一个似然方程如何变动。说明为什么在每层中所拟合的 XY 关联等于样本的相关系数。

9.40 当模型（XY,XZ,YZ）拟合不充分，且变量为定序变量时，有意义的模型嵌套于该模型和模型（XYZ）之间。对于定序的赋值 $\{u_i\}$、$\{v_j\}$ 以及 $\{w_k\}$，考虑

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \beta u_i v_j w_k。\quad (9.22)$$

a. 定义 $\theta_{ijk} = \theta_{ij(k+1)}/\theta_{ij(k)} = \theta_{i(j+1)k}/\theta_{i(j)k} = \theta_{(i+1)jk}/\theta_{(i)jk}$。对于单位距离的赋值,证明 $\log \theta_{ijk} = \beta$。Goodman(1979a)将此称为*一致性交互效应模型*(*uniform interaction model*)。

b. 证明任意两个变量之间的对数发生比之比在第三个变量的取值水平上呈线性变动。

c. 证明它的似然方程等于模型(*XY*,*XZ*,*YZ*)的似然方程再加上

$$\sum_i \sum_j \sum_k u_i v_j w_k \hat{\mu}_{ijk} = \sum_i \sum_j \sum_k u_i v_j w_k n_{ijk}。$$

d. 说明为什么式 9.12 模型是式 9.22 模型的一个特例。

9.41 构建一个包括一般的 *XZ* 和 *YZ* 关联项的模型,但关于 *XY* 关联的行效应为:(a)同质的,(b)在 *Z* 的不同取值间是异质的。加以解释。

9.42 说明为什么 *RC* 模型要求对赋值进行刻度限定。证明它的残差自由度为 $df = (I-2)(J-2)$。给出它的似然方程并加以解释。说明为什么拟合结果不会随着类别间排序的变化而变化。

9.43 参考式 9.16 相关模型(Goodman,1985,1986)。

a. 证明 λ 是赋值之间的相关系数。

b. 如果这个模型成立,证明 $\sum_i \mu_i(\pi_{ij}/\pi_{+j}) = \lambda v_j$ 并且 $\sum_j v_j(\pi_{ij}/\pi_{i+}) = \lambda \mu_i$。加以解释。

c. 当 λ 接近于零时,证明 $\log(\pi_{ij})$ 具有 $\gamma_i + \delta_i + \lambda \mu_i v_j + o(\lambda)$ 的形式,其中当 $\lambda \to 0$ 时 $o(\lambda)/\lambda \to 0$。因此,当这个关联很弱时,相关模型与 $\beta = \lambda$ 且赋值为 $\{u_i = \mu_i\}$ 和 $\{v_j = v_j\}$ 的双线性关联模型相似。

9.44 对于一般性的典型相关模型,证明 $\sum \lambda_k^2 = \sum_i \sum_j (\pi_{ij} - \pi_{i+}\pi_{+j})^2/\pi_{i+}\pi_{+j}$。因此,相关系数的平方分割成一系列相依指标,它们等于 $n=1$ 的独立性模型的 X^2 的非中心参数式 6.8(Goodman(1986)给出了其他的分割方法)。

9.45 参考式 9.18 模型。给定历险时间 $\{t_{ij}\}$,证明它的充分统计量是 $\{n_{i+}\}$ 和 $\{n_{+j}\}$。

9.46 参考第 9.7.3 节。令 $T = \sum t_i$,$W = \sum w_i$。假定存活时间服从参数为 λ 的负指数分布。

a. 利用式 9.19 对数似然函数,证明 $\hat{\lambda} = W/T$。

b. 以 T 为条件,证明 W 服从均值为 $T\lambda$ 的泊松分布。利用泊松分布似然函数,证明 $\hat{\lambda} = W/T$。

9.47 证明关于表 9.15 的最大似然估计不存在(提示:参见 Haberman(1973b,1974a,p. 398):如果 $\hat{\mu}_{111} = c > 0$,那么模型满足的边际限定条件意味着 $\hat{\mu}_{222} = -c$)。

9.48 对于对数线性模型,说明为什么在给定其他参数的充分统计量取值的情况下,当一个参数的充分统计量取其最小或最大的可能值时,对该参数的最大似然估计等于无穷大。

10 关于配对数据的模型

接下来,我们介绍比较两个匹配样本中的分类结果变量的方法,其中一个样本中的每个观测值与另一样本中的观测值两两配对。这样的*配对*(*matched-pairs*)数据一般发生在对研究对象进行重复测量时,比如,在不同时点对相同对象进行观察的*跟踪研究*(*longitudinal studies*)。由于这种匹配关系,两个样本中的结果变量在统计上并不独立。本书以下四章都是介绍处理此类相依数据的特定方法,本章是其中的第一章。

表 10.1 给出了配对数据的一个例子。在对英国一个 1 600 名适龄选民的随机样本进行的民意测验中,944 人表示支持首相的施政表现。六个月后,同样是这 1 600 人,表示支持的有 880 人。该表主对角线上的单元格行和列的选项都相同,这些调查对象在两次调查中表达了相同的观点。他们是样本的主体,因为只有相对较少的人在两次调查中改变了观点。调查对象在六个月前后的观点之间具有很强的关联,样本的发生比之比为 $(794 \times 570)/(150 \times 86) = 35.1$。

表 10.1 对首相施政表现的评价

第一次调查	第二次调查		小计
	支持	不支持	
支持	794	150	944
不支持	86	570	656
小计	880	720	1 600

对于分类变量的配对数据,可以通过行和列具有相同类别的二维列联表来表示,这是一个方形(*square*)表格。本章将介绍分析方形表格的方法。在第 10.1 节中,我们介绍比较二分结果变量的比例的方法。第 10.2 节讨论关于配对数据的 logistic 回归分析。第 10.3 节介绍结果变量具有多个类别的情况,包括相应的定类和定序 Logit 模型。第 10.4 节讨论关于方形表格的对数线性模型。在第 10.5 节和第 10.6 节中,我们讨论有关配对分析方法的两个应用,其中都用到关于方形表格的模型:一是分析两个评定者对同一组对象进行打分的一致性,二是基于成对的评估结果来评价对干预方式的偏好。

第 10.7 节将第 10.2 节至第 10.4 节介绍的模型扩展到多维表格的情况,这些多维表格是由观测值的匹配集形成的。我们将在第 11 章进一步介绍包括解释变量的情况。

10.1 相依比例的比较

对于数据中的 n 个配对,令 π_{ab} 表示每个配对中第一个观测值结果为 a 而第二个观测值结果为 b 的概率。令 n_{ab} 表示这类配对的计数,$p_{ab} = n_{ab}/n$ 就是相应的样本比例。我们将 $\{n_{ab}\}$ 视为服从多项分布 $(n;\{\pi_{ab}\})$ 的一个样本。这时,p_{a+} 是配对中第 1 个观测值落在类别 a 的比例,p_{+a} 是第 2 个观测值落在类别 a 的相应比例。我们通过比较边际比例 $\{p_{a+}\}$ 和 $\{p_{+a}\}$ 来对比这两个样本。在匹配样本中,这些比例是相关的,因而有关分析独立样本的方法在此并不适用。

在本节中,我们考虑结果变量为二分变量的情况。当 $\pi_{1+} = \pi_{+1}$ 时,就有 $\pi_{2+} = \pi_{+2}$,进而数据满足*边际同质性*(*marginal homogeneity*)。由于

$$\pi_{1+} - \pi_{+1} = (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21},$$

在 2 × 2 表格中,边际同质性等价于 $\pi_{12} = \pi_{21}$。这时,表格表现出关于主对角线的*对称性*(*symmetry*)。

10.1.1 相依比例的统计推断

比较边际分布的一个指标是 $\delta = \pi_{+1} - \pi_{1+}$。令

$$d = p_{+1} - p_{1+} = p_{2+} - p_{+2}。$$

根据多项分布的协方差公式式 1.3, $\text{cov}(p_{+1},p_{1+}) = \text{cov}(p_{11} + p_{21},p_{11} + p_{12})$ 可简化为 $(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/n$。因此,

$$\text{var}(\sqrt{n}d) = \pi_{1+}(1 - \pi_{1+}) + \pi_{+1}(1 - \pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})。\quad (10.1)$$

在大样本的情况下,d 近似服从正态样本分布。这时,关于 $\delta = \pi_{+1} - \pi_{1+}$ 的置信区间为

$$d \pm z_{\alpha/2}\hat{\sigma}(d),$$

其中

$$\hat{\sigma}^2(d) = [p_1 + (1 - p_{1+}) + p_{+1}(1 - p_{+1}) - 2(p_{11}p_{12} - p_{12}p_{21})]/n$$
$$= [(p_{12} + p_{12}) - (p_{12} - p_{21})^2]/n, \quad (10.2)$$

对上述第一个等式进行必要的代数运算便可得出第二个等式。对 $H_0:\delta = \delta_0$ 的计分检验进行逆运算更加复杂,但是它所给出的置信区间的涵盖概率更接近于其名义置信水平(Tango,1998)。在计算 d 和 $\hat{\sigma}(d)$ 之前,先对每个单元格计数加 1 也具有同样的效果。

边际同质性假设为 $H_0:\pi_{1+} = \pi_{+1}$(即,$\delta = 0$)。沃尔德统计量为 $z = d/\hat{\sigma}(d)$ 或其平方。在 H_0 下,关于方差的另一个估计为

$$\hat{\sigma}_0^2(d) = \frac{p_{12} + p_{21}}{n} = \frac{n_{12} + n_{21}}{n^2}。\quad (10.3)$$

计分检验统计量 $z_0 = d/\hat{\sigma}_0(d)$ 可简化为

$$z_0 = \frac{n_{21} - n_{12}}{(n_{21} + n_{12})^{1/2}}。\quad (10.4)$$

z_0 的平方是 df = 1 的卡方统计量。利用该统计量进行的检验被称为 *McNemar 检验*(*McNemar's test*)(McNemar,1947)。

McNemar 统计量仅取决于同一配对中两个观测值取值*不同*的情况。主对角线上的 $n_{11} + n_{22}$ 与推断 π_{1+} 是否不同于 π_{+1} 无关。这个结论可能有些出人意料,但实际上*所有的*

配对都会影响关于 π_{1+} 和 π_{+1} 到底*相差多少*的统计推断,例如,δ 及其标准误的估计。

10.1.2 例子:首相支持率

在表 10.1 中,第一次民意测验支持首相表现的样本比例为 $p_{1+} = 944/1\ 600 = 0.59$,第二次为 $p_{+1} = 880/1\ 600 = 0.55$。根据式 10.2,$\pi_{+1} - \pi_{1+}$ 的 95% 的置信区间等于 $(0.55 - 0.59) \pm 1.96(0.009\ 5)$,即 $(-0.06, -0.02)$。支持率在此期间似乎下降了 2% 到 6%。

在检验边际同质性时,利用零分布方差的检验统计量式 10.4 等于

$$z_0 = \frac{86 - 150}{(86 + 150)^{1/2}} = -4.17。$$

该检验给出了支持率下降的明显证据。

10.1.3 利用相依样本提高推断精度

公式 10.1 的最后一项与 $\text{cov}(p_{+1}, p_{1+})$ 有关,它反映了边际比例之间的相依性。与之相对照,在样本规模分别为 n 的独立样本中估计二项分布概率 π_1 和 π_2 时,样本比例之间的协方差等于零,并且

$$\text{var}[\sqrt{n}(\text{样本比例之差})] = \pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)。$$

相依样本之间往往会展示出一种正向的相依关系,满足 $\log\theta = \log[\pi_{11}\pi_{22}/\pi_{12}\pi_{21}] > 0$,也即,$\pi_{11}\pi_{22} > \pi_{12}\pi_{21}$。由式 10.1 可得,正向相依意味着 $\text{var}(d)$ 比独立样本时的值小。

在研究设计中使用相依样本有助于提高关于对象内效应(within-subject effects)统计推断的精度(相反,在给定观测值数量的情况下,进行对象间组别比较(between-subject group comparison)的标准误一般会偏大)。当样本间高度相关时,这种精度的提高非常明显。举例来说,表 10.1 中的相依样本的样本规模均为 1 600,这时,$d = 0.55 - 0.59$ 的标准误为 0.009 5。这两组观测值之间存在强关联,样本的发生比之比等于 35.1。相比之下,在两个具有相同规模的独立样本中,$\hat{\pi}_1 - \hat{\pi}_2 = 0.55 - 0.59$ 的标准误等于 0.017 5,几乎是前者的两倍。

10.1.4 比较匹配比例的小样本检验

配对数据中二分变量的边际同质性零假设等价于 $H_0: \pi_{12} = \pi_{21}$ 或 $\pi_{21}/(\pi_{21} + \pi_{12}) = 0.5$。在小样本的情况下,精确检验以 $n^* = n_{21} + n_{12}$ 为条件(Mosteller,1952)。在 H_0 下,n_{21} 服从参数为 $\left(n^*, \frac{1}{2}\right)$ 的二项分布,并有 $E(n_{21}) = \frac{1}{2}n^*$。精确检验的 P 值等于二项分布的尾部概率。

举例来说,在表 10.1 中,考虑 $H_a: \pi_{+1} < \pi_{1+}$,或等价地,$H_a: \pi_{21} < \pi_{12}$。由于 $n^* = 86 + 150 = 236$,检验的参照分布为 $\text{bin}(236, \frac{1}{2})$。检验的 P 值等于在 236 次试验中至少成功 150 次的概率,即为 0.000 02。关于 $H_a: \pi_{+1} \neq \pi_{1+}$ 的 P 值是该值的两倍。

当 $n^* > 10$ 时,所参照的二项分布近似于均值为 $\frac{1}{2}n^*$ 和方差为 $n^*\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ 的正态分布。标准化的正态检验统计量等于

$$z = \frac{n_{21} - \frac{1}{2}n^*}{\left[n^*\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\right]^{1/2}} = \frac{n_{21} - n_{12}}{(n_{21} + n_{12})^{1/2}}。$$

这与 McNemar 统计量式 10.4 相同。

10.1.5 McNemar 检验与 Cochran-Mantel-Haenszel 检验之间的联系

二分结果变量的 n 个配对数据可以通过 n 个分表来表示,其中每个配对对应于一个 2×2 表格。分表的列代表每次测量的两种可能结果,第 1 行表示配对中第一个观测值的结果,而第 2 行表示第二个观测值的结果。

表 10.2 给出了这种方式对应的四种可能的分表。表 10.1 的完整三维表格共包括 1 600个这样的分表,其中 794 个具有表 10.2 中第 1 个对象的形式(即,在两次调查中都"支持"),570 个在两次调查中都"不支持"者具有第 2 个对象的分表形式,86 个具有第 3 个对象的形式,以及 150 个具有第 4 个对象的形式。在 $2\times2\times1\ 600$ 的列联表中,表10.1 的 1 600 个研究对象共提供了 3 200 个观测值。将 1 600 个分表进行合并得到一个 2×2 表格,其中第一行等于 (944,656),第二行等于 (880,720)。这些是在两次调查中(支持,不支持)的总数,组成了表 10.1 中的边际总计。

表 10.2 表 10.1 所包括的四种配对形式

调查对象	调查时间	结果	
		支持	不支持
1	第一次	1	0
	第二次	1	0
2	第一次	0	1
	第二次	0	1
3	第一次	0	1
	第二次	1	0
4	第一次	1	0
	第二次	0	1

就每个调查对象来说,假定在每次调查中表示支持的概率是相等的。那么,在控制调查对象后,调查结果与调查时间之间条件独立。这时,在对所有对象进行合并后所形成的边际表中,每次调查的支持率也一样。但是,这意味着表 10.1 的真实概率满足边际同质性。因此,对 $2\times2\times1\ 600$ 表格进行的条件独立性检验同时也是关于表 10.1 的边际同质性检验。

我们可以使用 Cochran-Mantel-Haenszel(CMH)统计量(式 6.6)来对这个三维表格进行条件独立性检验。这个卡方统计量在代数上等于 McNemar 统计量的平方,对于形如表 10.1 的表格,该统计量为 $(n_{21}-n_{12})^2/(n_{12}+n_{21})$。McNemar 检验是 CMH 检验的一个特例,适用于由 n 个分表表示的关于二分结果变量的 n 个配对数据。从运算的角度来说,这种联系没有什么实际意义,因为 McNemar 统计量的计算很简便。不过,它确实表明了处理更复杂的配对数据的方法。在结果变量存在多个类别或匹配集包括多个观测值的情况下,可以利用广义 CMH 检验(第 7.5 节)来检验边际同质性,将每个对象视为一个层,每行表示一个特定的观测值(Darroch,1981;Mantel and Byar,1978)。

在接下来的几节中,我们将关于配对数据的 $2\times2\times n$ 表格称为*对象别*(*subject-specific*)表格,并将形如表 10.1 的 2×2 表格称为*总体平均*(*population-averaged*)表格,因为后者的边际提供了关于总体中边际比例的直接估计。

10.2 二分配对数据的条件 Logistic 回归

在第 6.7 节中,我们介绍了从分析中剔除冗余参数的*条件 logistic 回归*(*conditional logistic regression*)。现在,我们讨论如何利用该模型分析二分的配对数据。该模型对应的是对象别表格。

10.2.1 配对数据的边际模型与条件模型

第 10.1 节的分析可以在模型的框架下完成。令 (Y_1, Y_2) 表示对一个随机选取的研究对象的两次观测值,其中结果"1"代表结果变量的第一个类别(成功),结果"0"代表第二个类别。两个边际概率之差 $\delta = P(Y_2 = 1) - P(Y_1 = 1)$ 是以下模型的参数:

$$P(Y_t = 1) = \alpha + \delta x_t, \tag{10.5}$$

其中 $x_1 = 0$ 且 $x_2 = 1$,这时,$P(Y_1 = 1) = \alpha$, $P(Y_2 = 1) = \alpha + \delta$。相应地,使用 Logit 连结的模型可表示为

$$\text{Logit}[P(Y_t = 1)] = \alpha + \beta x_t。 \tag{10.6}$$

参数 β 是边际分布之间的对数发生比之比。

式 10.5 和式 10.6 模型都属于*边际模型*(*marginal models*):这些模型关注的是两次观测中结果变量的边际分布。例如,在式 10.6 模型中,关于 β 的最大似然估计等于总体平均表格中边际比例的对数发生比之比,即 $\hat{\beta} = \log[p_{+1}p_{2+}/p_{+2}p_{1+}]$。相应的渐近方差,参见习题 10.26。

与之相对照,类似表 10.2 的对象别表格允许结果发生的概率在不同调查对象之间存在差异。令 (Y_{i1}, Y_{i2}) 表示第 i 对观测值,$i = 1, \cdots, n$。这时,模型的形式为

$$连结[P(Y_{it} = 1)] = \alpha_i + \beta x_t。 \tag{10.7}$$

此模型被称为*条件模型*(*conditional model*),因为效应 β 的定义是以对象本身为条件的。它所估计的是以调查对象作为层的三维表格的条件关联。这个效应是*对象别*(*subject-specific*)效应,它是在对象层面上界定的。与之相对,式 10.5 和式 10.6 边际模型中的效应是*总体平均*(*population-averaged*)效应,因为它们是针对整个总体进行平均而不是针对个体对象。

在恒等连结的情况下,对象别效应与总体平均效应是一致的。例如,在使用恒定连结的条件模型式 10.7 中,对于所有 i 都有 $\beta = P(Y_{i2} = 1) - P(Y_{i1} = 1)$,将总体中的对象求平均值可得 β 等于式 10.5 模型中的参数 δ。但是,在非线性连结的情况下,对象别效应并不等于总体平均效应。例如,对于使用 Logit 连结的式 10.7 模型,

$$P(Y_{it} = 1) = \frac{\exp(\alpha_i + \beta x_t)}{[1 + \exp(\alpha_i + \beta x_t)]}。$$

在总体中对此求平均值,并不对应于式 10.6 边际 Logit 模型中的 $\exp(\alpha + \beta x_t)/[1 + \exp(\alpha + \beta x_t)]$。下面,我们重点介绍使用 Logit 连结的条件模型。

10.2.2 关于对象别概率的 Logit 模型

式 10.7 模型不同于前面章节中允许每个调查对象都有自己的概率分布的模型。Cox(1958b,1970)以及 Rasch(1961)介绍了对该模型使用 Logit 连结的情况。令 Y_{it} 表示第 i 个调查对象的第 t 次观测值,这个模型可表示为

$$\text{Logit}[P(Y_{it}=1)] = \alpha_i + \beta x_t, \tag{10.8}$$

其中 $x_1 = 0$ 且 $x_2 = 1$。尽管该模型允许不同的对象别分布，但是它假定效应 β 是相同的。对于第 i 个对象，

$$P(Y_{i1}=1) = \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}, \qquad P(Y_{i2}=1) = \frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}。$$

参数 β 用来比较结果变量的分布。对于每个对象，第 2 个观测值的结果为成功的发生比是第 1 个观测值的 $\exp(\beta)$ 倍。

给定这些参数，式 10.8 模型一般假定不同对象之间以及同一对象的两个观测值之间结果变量的取值是相互独立的。然而，当对所有对象进行平均时，结果变量的取值存在非负的关联关系。假定 $|\beta|$ 比 $|\alpha_i|$ 小很多，那么，对于 α_i 取一个很大的正值的对象，每次观测中 $P(Y_{it}=1)$ 都较大，因而很可能观测到的结果为成功；对于 α_i 取绝对值很大的负值的对象，每次观测中 $P(Y_{it}=1)$ 都较小，因而很可能观测到的结果为失败。$\{\alpha_i\}$ 的变动性越大，结果变量取值之间的总体正向关联也就越强，即当第 1 个观测值结果为成功（失败）时，第 2 个观测值也倾向于为成功（失败）。这一点对于任意的 β 都适用。正向关联反映了每一配对中的两个观测值所共同具有的 α_i 的影响。只有当所有 $\{\alpha_i\}$ 都相等时，才不存在上述关联。因此，这个模型确实体现了配对数据的相依性。通过模型本身的结构，对它的拟合考虑了这种非负关联。

在这个模型中，当 $\{\alpha_i\}$ 的取值很大时，拟合过程可能会出现问题，并影响普通最大似然估计值所具有的那些特性（习题 10.24）。相应地，条件最大似然法对此的解决办法是，将 $\{\alpha_i\}$ 作为冗余参数，对消除它们后的条件分布的似然函数进行最大化。在术语上需要注意的是：我们将式 10.8 模型称为*条件模型*，是强调该模型中的效应 β 是对象别的，以对象本身为条件。下文将要介绍的关于这种模型的分析属于*条件* logistic 回归的例子，但是在这里，*术语条件*（*conditional*）指的是以冗余参数的充分统计量为条件，在最大似然分析中消除相应的冗余参数。

10.2.3 二分配对数据的条件最大似然推断

在式 10.8 模型中，假定不同对象之间以及同一对象的两个观测值之间结果变量的取值相互独立，关于 $\{(y_{11},y_{12}),\cdots,(y_{n1},y_{n2})\}$ 的联合密度函数为

$$\prod_{i=1}^{n}\left(\frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\right)^{y_{i1}}\left(\frac{1}{1+\exp(\alpha_i)}\right)^{1-y_{i1}} \times$$
$$\left(\frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}\right)^{y_{i2}}\left(\frac{1}{1+\exp(\alpha_i+\beta)}\right)^{1-y_{i2}}。$$

从数据的角度来说，上式与

$$\exp\left[\sum_i \alpha_i(y_{i1}+y_{i2}) + \beta\left(\sum_i y_{i2}\right)\right]$$

成比例。为了消除 $\{\alpha_i\}$，我们以 $\{\alpha_i\}$ 的充分统计量为条件，即成对的成功总计 $\{S_i = y_{i1} + y_{i2}\}$。给定 $S_i = 0$，$P(Y_{i1} = Y_{i2} = 0) = 1$；给定 $S_i = 2$，$P(Y_{i1} = Y_{i2} = 1) = 1$。仅当 $S_i = 1$，也即仅当两个结果变量的取值不同时，(Y_{i1}, Y_{i2}) 的分布才取决于 β。给定 $y_{i1} + y_{i2} = 1$，相应的条件分布等于

$$\begin{aligned} &p(Y_{i1}=y_{i1}, Y_{i2}=y_{i2} \mid S_i = 1) \\ &\quad = P(Y_{i1}=y_{i1}, Y_{i2}=y_{i2}) / [P(Y_{i1}=1, Y_{i2}=0) + P(Y_{i1}=0, Y_{i2}=1)] \end{aligned}$$

$$= \frac{\left(\frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\right)^{y_{i1}}\left(\frac{1}{1+\exp(\alpha_i)}\right)^{1-y_{i1}}\left(\frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}\right)^{y_{i2}}\left(\frac{1}{1+\exp(\alpha_i+\beta)}\right)^{1-y_{i2}}}{\frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\frac{1}{1+\exp(\alpha_i+\beta)}+\frac{1}{1+\exp(\alpha_i)}\frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}}$$

$$= \frac{\exp(\beta)}{[1+\exp(\beta)]}(\text{当 } y_{i1}=0, y_{i2}=1)$$

$$= \frac{1}{[1+\exp(\beta)]}(\text{当 } y_{i1}=1, y_{i2}=0)。$$

再次,令 $\{n_{ab}\}$ 表示四种可能情况对应的计数。对于 $S_i = 1$ 的调查对象,$\sum_i y_{i1} = n_{12}$,即第1个观测值为成功且第2个观测值为失败的对象的数量。类似地,有 $\sum_i y_{i2} = n_{21}$,并且 $\sum_i S_i = n^* = n_{12} + n_{21}$。由于 n_{21} 是对 n^* 个独立同分布的伯努利变量的求和,它的条件分布是参数为 $\exp(\beta)/[1+\exp(\beta)]$ 的二项分布。在检验边际同质性($\beta = 0$)时,该参数等于 $\frac{1}{2}$。总之,关于 Logit 模型的条件分析意味着,具有 $y_{i1} = y_{i2}$ 的配对与对 β 的统计推断无关。当这个模型在现实应用中成立时,它为在比较边际分布时仅使用两个观测值取值不同的 $n_{12} + n_{21}$ 个配对数据提供了依据。

以 $S_i = 1$ 为条件,配对数据的联合分布为

$$\prod_{S_i=1}\left(\frac{1}{1+\exp(\beta)}\right)^{y_{i1}}\left(\frac{\exp(\beta)}{1+\exp(\beta)}\right)^{y_{i2}} = \frac{[\exp(\beta)]^{n_{21}}}{[1+\exp(\beta)]^{n^*}}, \tag{10.9}$$

其中乘积项包括了所有满足 $S_i = 1$ 的配对。对这个条件似然函数的对数求导,令其导数等于0,并对该方程求解便得到了式 10.8 模型中关于 β 的条件最大似然估计。读者可以自行验证,这个估计值及其标准误分别等于

$$\hat{\beta} = \log(n_{21}/n_{12}), \text{SE} = \sqrt{1/n_{21} + 1/n_{12}}。 \tag{10.10}$$

10.2.4 二分配对数据模型中的随机效应

处理式 10.8Logit 模型中的大量冗余参数 $\{\alpha_i\}$ 的另一种方法是将其作为*随机效应*(*random effects*)。这种方法将 $\{\alpha_i\}$ 视为服从特定概率分布的一个未观测到的随机样本,通常假定 $\{\alpha_i\}$ 服从 $N(\mu,\sigma^2)$ 分布,其中 μ 和 σ 为未知参数。该方法按照 $\{\alpha_i\}$ 的分布,通过对其求平均来消除这些参数,从而得到相应的边际分布。这时,似然方程取决于 β 以及 $N(\mu,\sigma^2)$ 的参数。它总共只有三个参数,因而更易于处理。当配对数据存在非负的样本对数发生比之比时,这种方法同样得出 $\hat{\beta} = \log(n_{21}/n_{12})$(Neuhaus et al. ,1994)。这个模型属于*广义线性混合模型*(*generalized linear mixed model*),它既包括随机效应,又包括固定效应 β。该模型的详细内容,我们将在第 12 章介绍。

式 10.8 模型隐含着,n 个对象别分表中的所有真实发生比之比都等于 $\exp(\beta)$。在第 6.3.5 节中,我们介绍了关于多个 2 × 2 表格的共同发生比之比的 Mantel-Haenszel 估计。实际上,表 10.2 所示的对象别表格的该估计值在代数上与表 10.1 所示数据形式对应的 n_{21}/n_{12} 相等(回顾结果变量所有取值都只落在一列的分表对 CMH 检验或 Mantel-Haenszel 估计没有意义)。总之,Mantel-Haenszel 估计、条件最大似然估计以及对包括随机效应的式 10.8 Logit 模型的最大似然估计(具有非负的对数发生比之比)都得出 $\exp(\hat{\beta}) = n_{21}/n_{12}$。

10.2.5 关于匹配的个案—控制研究的 Logistic 回归

配对数据中的两个观测值 (y_{i1}, y_{i2}) 并不一定都是对同一对象的两次测量。例如，在个案—控制研究中，对每个个案匹配一个控制案例也会形成配对数据。当结果变量 Y 为二分变量时，按照可能影响结果变量的一组条件，对每个个案 $(Y=1)$ 匹配一个控制案例 $(Y=0)$。对配对中的对象测量预测变量 X 的值，然后分析 XY 的关联。

表 10.3 展示了一项个案—控制研究。该研究将患有急性心肌梗塞的 144 名纳瓦霍印第安人(Navajo Indians)按照年龄和性别与 144 名未患心脏病的人进行了匹配。这些研究对象被问及是否曾经被诊断出过糖尿病($x=0$, 否; $x=1$, 是)。表 10.3 具有与表 10.1 相同的形式，只是行和列由 X 的类别(而不是 Y 的类别)构成。

匹配的个案—控制数据也可以表示成形如表 10.2 的分表形式，但这时需要互换 X 和 Y 的角色。如表 10.4 所示，X 的取值存在四种可能的模式。因为数据中共有 37 个配对个案患过糖尿病而控制案例未患过糖尿病，所以 a 类型的分表有 37 个。相应地，数据包括 16 个 b 类型的分表，9 个 c 类型的分表以及 82 个 d 类型的分表。

现在，针对第 i 个配对中的第 t 个对象，考虑模型

$$\text{Logit}\left[P(Y_{it}=1)\right] = \alpha_i + \beta x_{it}。 \tag{10.11}$$

该模型中的概率是指给定 X 后 Y 的分布，然而，回顾性研究只提供给定 Y 后 X 的分布情况。不过，我们仍可以估计发生比之比 $\exp(\beta)$，因为它指的是 XY 的发生比之比，与两个条件分布都有关(第 2.2.4 节，第 5.1.4 节)。尽管这项研究在哪个是固定变量、哪个是随机变量方面互换了 X 和 Y 的角色，关于 $\exp(\beta)$ 的条件最大似然估计仍为 $n_{21}/n_{12}=37/16=2.3$。

表 10.3 关于心肌梗塞的个案—控制配对数据中曾患糖尿病的情况

心肌梗塞控制案例	心肌梗塞个案		小 计
	曾患糖尿病	未患糖尿病	
曾患糖尿病	9	16	25
未患糖尿病	37	82	119
小计	46	98	144

来源：J. L. Coulehan et al., *Amer. J. Public Health* 76:412-414(1986)，经美国公共卫生学会(the American Public Health Association)授权重印。

表 10.4 表 10.3 中个案—控制配对的可能类型

糖尿病	a		b		c		d	
	个案	控制	个案	控制	个案	控制	个案	控 制
是	1	0	0	1	1	1	0	0
否	0	1	1	0	0	0	1	1

10.2.6 包括多个预测变量的配对数据的条件最大似然估计

当关于二分结果变量的个案—控制或对象别配对数据中包括 p 个预测变量时，模型一般化为

$$\text{Logit}[P(Y_{it}=1)] = \alpha_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_p x_{pit}, \tag{10.12}$$

其中 x_{hit} 表示在第 i 个配对中第 t 个观测值对应的第 h 个预测变量的取值，$t=1,2$。一般

来说,其中某一预测变量是所关注的解释变量,如糖尿病状况,其他的预测变量是除了那些在形成配对时已经控制的变量之外仍需要控制的协变量。可以通过条件最大似然法来估计 $\{\beta_j\}$,以 α_i 的充分统计量为条件,从而在似然函数中消除这些参数。

令 $\mathbf{x}_{it} = (x_{1it},\cdots,x_{pit})'$ 以及 $\boldsymbol{\beta} = (\beta_1,\cdots,\beta_p)'$,对第 10.2.3 节的推导加以扩展可得:

$$P(Y_{i1} = 0, Y_{i2} = 1 \mid S_i = 1) = \frac{\exp(\mathbf{x}'_{i2}\boldsymbol{\beta})}{\left[\exp(\mathbf{x}'_{i1}\boldsymbol{\beta}) + \exp(\mathbf{x}'_{i2}\boldsymbol{\beta})\right]},$$

$$P(Y_{i1} = 1, Y_{i2} = 0 \mid S_i = 1) = \frac{\exp(\mathbf{x}'_{i1}\boldsymbol{\beta})}{\left[\exp(\mathbf{x}'_{i1}\boldsymbol{\beta}) + \exp(\mathbf{x}'_{i2}\boldsymbol{\beta})\right]}。 \quad (10.13)$$

将分子和分母都除以 $\exp(\mathbf{x}'_{i1}\boldsymbol{\beta})$ 表明,第一个方程具有不包括截距项的 logistic 回归的形式,其预测项的值为 $\mathbf{x}_i^* = \mathbf{x}_{i2} - \mathbf{x}_{i1}$。事实上,也可以通过对这些配对单独拟合 logistic 回归模型来求得关于式 10.12 模型的条件最大似然估计。相应模型的结果变量为 y^*,当 $(y_{i1} = 0, y_{i2} = 1)$ 时 $y^* = 1$,当 $(y_{i1} = 1, y_{i2} = 0)$ 时 $y^* = 0$,模型不包括截距项,预测变量为 $\mathbf{x}_i^*$。这个模型的似然函数与条件似然函数相同(Breslow et al.,1978;Chamberlain,1980)。

举例来说,对表 10.3 的数据拟合式 10.11 模型,令 $y_i^* = y_{i2} - y_{i1}$ 以及 $x_i^* = x_{i2} - x_{i1}$。如果 $t = 1$ 表示控制案例、$t = 2$ 表示个案,那么总存在 $y_i^* = 1$。由于 $x_{it} = 1$ 代表患过糖尿病,$x_{it} = 0$ 表示没有,16 个观测值具有 $(y_i^* = 1, x_i^* = -1)$,$9 + 82 = 91$ 个观测值具有 $(y_i^* = 1, x_i^* = 0)$,另外 37 个观测值具有 $(y_i^* = 1, x_i^* = +1)$。设定 $\hat{\alpha} = 0$ 的 Logit 模型的拟合结果为 $\hat{\beta} = 0.84$。在只包括一个二分预测变量的情况下,该估计值等于 $\log(n_{21}/n_{12})$。

10.2.7 对边际模型和条件模型的扩展

对于二分的配对数据,我们在第 10.1 节介绍了边际(即总体平均)模型,本节又介绍了相应的条件(即对象别)模型。这些模型可以扩展到结果变量为多项分布变量以及数据为匹配集(matched sets)的情况。例如,Chamberlain(1980)讨论了多项分布结果变量配对数据的条件最大似然估计。对于二分的结果变量,式 10.12 模型适用于 α_i 是一组关于对象 i 的重复测量的情况。或者,α_i 也可以是一个匹配集,即由多个对象形成的群组(cluster),如第 i 个家庭中的孩子或第 i 窝中的胚胎。

在将条件模型扩展到匹配集的群组时,条件最大似然法所估计的 β_j 被限定为组内效应(within-cluster effects),如个案—控制研究和交叉研究(crossover studies)中所出现的。这种情况下,对于每个 i,解释变量的取值随着 t 的变动而变动。条件最大似然法无法估计组间效应(between-cluster effects)。关于组间效应的统计量需要利用在相应的解释变量不同取值水平上的研究对象的总计,但是,这些总计是 $\{\alpha_i\}$ 的充分统计量之和,因而它们本身是固定的,并且在以 $\{\alpha_i\}$ 的充分统计量为条件后具有退化的分布(degenerate distributions)。条件似然函数会消除掉在每个 i 中对所有 t 取值都一样的解释变量(在式 10.13 的配对数据中,对于任意 j,如果所有 i 个对象都有 $x_{ji1} = x_{ji2}$,就会出现这种情况)。对于该变量,我们最多可以按照其取值对数据进行分层,然后对每一层分别拟合估计组内效应的模型。随机效应方法相对于条件最大似然法的一个优点在于,它可以不仅仅局限于估计组内效应。

在本章余下的内容中,我们将重点介绍有关多项分布结果变量的配对数据的边际模型。在下一章中,我们将边际模型扩展到允许匹配集以及包括解释变量的情形。通过随

机效应方法拟合条件模型在运算上尤其复杂。在本章我们会简要提及一些多项分布条件模型,相应内容的详细讨论将留到第 12 章。

10.3　方形列联表的边际模型

配对数据的分析可以从二分变量扩展到结果变量具有 $I > 2$ 个类别的情况。$I \times I$ 的方形表 $\{n_{ab}\}$ 给出了关于 (Y_1, Y_2) 的可能结果 (a,b) 的计数,令 $\pi_{ab} = P(Y_1 = a, Y_2 = b)$。边际同质性意味着对于所有 $a = 1,\cdots,I$, 存在 $P(Y_1 = a) = P(Y_2 = a)$。边际模型旨在比较概率 $\{P(Y_1 = a)\}$ 和 $\{P(Y_2 = a)\}$。

10.3.1　定序测度的边际模型

当类别间存在排序时,关于二分配对数据的边际模型式 10.6 可以通过定序 Logit 进行扩展。在累积 Logit 的情况下,

$$\text{Logit}[P(Y_t \leqslant j)] = \alpha_j + \beta x_t, \quad t = 1,2, \quad j = 1,\cdots,I-1, \qquad (10.14)$$

其中 $x_1 = 0$ 且 $x_2 = 1$。这个模型具有比例发生比的结构(第 7.2.2 节)。结果 $Y_2 \leqslant j$ 的发生比等于结果 $Y_1 \leqslant j$ 的发生比的 $\exp(\beta)$ 倍。该模型隐含着相应的边际分布具有随机排序特征,$\beta > 0$ 时表明 Y_1 一般大于 Y_2。边际同质性相当于 $\beta = 0$。

在拟合上述模型时,将 (Y_1, Y_2) 视为相依的。最大似然法对 $\{\pi_{ab}\}$ 的多项分布似然函数进行最大化。这在运算上并不容易。由于模型涉及边际概率 $\{P(Y_1 = a) = \pi_{a+}\}$ 和 $\{P(Y_2 = b) = \pi_{+b}\}$,所以无法直接将模型公式代入关于联合分布概率的对数似然函数的核函数 $\sum_a \sum_b n_{ab} \log \pi_{ab}$。有关边际模型的最大似然拟合,我们将推后到第 11.2.5 节讨论。式 10.14 模型通过 I 个参数描述了 $2(I-1)$ 个边际概率,因而拟合检验的自由度为 $\text{df} = I - 2$。另外,也可以利用一些综合指标来进行边际分布的比较,比如对有关类别赋值后的均值之差(习题 10.38)。

10.3.2　例子:婚前性行为与婚外性行为

参见表 10.5。在一次综合社会调查中,被访者回答了他们关于婚前性行为(一对夫妇在结婚之前发生性行为)以及婚外性行为(一个已婚者与配偶以外的其他人发生性行为)的看法。回答类别包括:1 = 总是错的,2 = 几乎总是错的,3 = 有时候是错的,4 = 根本没错。

表 10.5　对婚前性行为和婚外性行为的态度

婚前性行为	婚外性行为				小计
	1	2	3	4	
1	144	2	0	0	146
2	33	4	2	0	39
3	84	14	6	1	105
4	126	29	25	5	185
小计	387	49	33	6	475

来源:1989 年综合社会调查,美国民意研究中心。

婚前性行为的样本累积边际比例为 (0.307,0.389,0.611),婚外性行为的相应比例为 (0.815,0.918,0.987)。这表明,在定序尺度上,关于婚前性行为的回答要高于关于婚

外性行为的回答。如果对这些类别使用赋值（1,2,3,4），婚前性行为的均值为2.69，与类别“有时候是错的”的赋值最接近；而关于婚外性行为的均值为1.28，与“总是错的”的赋值最为接近。

式10.14累积Logit模型的拟合结果为$\hat{\beta} = 2.51$（SE=0.13）。因而，存在很强的证据表明，总体人群对婚前性行为的认同程度比婚外性行为更高。边际同质性模型对应的$G^2 = 348.1$（df=3），而式10.14模型的拟合结果对应的$G^2 = 35.1$（df=2）。定序模型对数据拟合得并不是很好，但却大大优于边际同质性模型的拟合。在第10.4.7节中，我们将考虑拟合结果更好的模型。

10.3.3 定类测度的边际模型

在结果变量为定类变量的情况下，对每个Logit假定具有相同的效应是不合理的。基线类别Logit模型的形式为

$$\log[P(Y_t = j)/P(Y_t = I)] = \alpha_j + \beta_j x_t, \quad t = 1,2, j = 1,\cdots,I-1, \quad (10.15)$$

其中$x_1 = 0$，$x_2 = 1$。这个模型用$2(I-1)$个参数来表示$2(I-1)$个边际概率。因此，它是一个饱和模型。

边际同质性是上述模型中$\beta_1 = \cdots = \beta_{I-1} = 0$时的一个特例。为了拟合该模型，Lipsitz等(1990)以及Madanksy(1963)在这些约束条件下对关于$\{n_{ab}\}$的多项分布似然函数进行了最大化，通过迭代法可以得到拟合值$\{\hat{\mu}_{ab}\}$。利用G^2或X^2将模型拟合值与$\{n_{ab}\}$进行比较，就是相应的边际同质性检验，自由度为df=$I-1$。

Bhapkar(1966)利用边际比例的渐近正态性进行了边际同质性检验。令$d_a = p_{+a} - p_{a+}$，并令$\mathbf{d}' = (d_1,\cdots,d_{I-1})$。由于$\sum d_a = 0$，$d_I$是冗余的。关于$\sqrt{n}\mathbf{d}$的样本协方差矩阵$\hat{\mathbf{V}}$具有如下元素：

$$\text{对于 } a \neq b, \hat{v}_{ab} = -(p_{ab} + p_{ba}) - (p_{+a} - p_{a+})(p_{+b} - p_{b+}),$$

$$\hat{v}_{aa} = p_{+a} + p_{a+} - 2p_{aa} - (p_{+a} - p_{a+})^2。$$

现在，$\sqrt{n}[\mathbf{d} - E(\mathbf{d})]$渐近服从多元正态分布，所估计的协方差矩阵为$\hat{\mathbf{V}}$。在边际同质性成立时，$E(\mathbf{d}) = 0$，并且

$$W = n\mathbf{d}'\hat{\mathbf{V}}^{-1}\mathbf{d} \quad (10.16)$$

渐近服从df=$I-1$的卡方分布。该统计量是当式10.15模型使用恒等连结时关于模型参数的沃尔德检验统计量。Stuart(1955)指出，$W_0 = n\mathbf{d}'\hat{\mathbf{V}}_0^{-1}\mathbf{d}$是相应模型的计分检验统计量，它使用了样本零分布(*null*)的协方差矩阵$\hat{\mathbf{V}}_0$，其元素为

$$\text{对于 } a \neq b,\ \hat{v}_{ab0} = -(p_{ab} + p_{ba}),$$

$$\hat{v}_{aa0} = p_{+a} + p_{a+} - 2p_{aa}。$$

Ireland等(1969)指出，$W = W_0/(1 - W_0/n)$。当$I = 2$时，W_0是McNemar统计量，即式10.4的平方。

这些检验使用了比较I对边际比例的所有$I-1$个自由度。在定序变量的情况下，当类别数量较多并且各类别之间具有很强的相关性时，定序检验（其自由度为df=1）的统计效能要强得多(Agresti,1984,p.209)。

10.3.4 例子：迁移的情况

表10.6比较了某一美国居民的样本在1985年和1980年的居住地状况。在这五年间相对很少的人改变了居住地，95%的观测值落在了表格的主对角线上。边际同质性模

型的最大似然拟合结果给出 $G^2 = 240.8$ (df = 3)，如表 10.6 所示。有关样本边际比例之差的统计量给出的结果与此相似，例如，Bhapkar 统计量式 10.16 为 $W = 236.5$ (df = 3)。

表 10.6　1980—1985 年间的迁移情况以及边际同质性模型的拟合结果

1980 年的居住地	1985 年的居住地				小计
	东北部	中西部	南部	西部	
东北部	11 607	100	366	124	12 197
	(11 607)	(98.1)	(265 7)	(94.0)	(12 064.7)
中西部	87	13 677	515	302	14 581
	(88.7)	(13 677)	(379.1)	(323.3)	(14 377.1)
南部	172	255	17 189	270	18 486
	(276.5)	(350.8)	(17 819)	(287.3)	(18 733.5)
西部	63	176	286	10 192	10 717
	(92.5)	(251.3)	(269.8)	(10 192)	(10 805.6)
小计	11 929	14 178	18 986	10 888	55 981
	(12 064.7)	(14 377.1)	(18 733.5)	(10 805.6)	

来源：数据取自：Table 12 of U. S. Bureau of the Census, Current Population Reports, Series P-20, No. 420, *Geographical Mobility: 1985* (Washington, DC: U. S. Government Printing Office), 1987。

在 1980 年和 1985 年，四个地区的样本边际比例分别为 (0.218, 0.260, 0.330, 0.191) 和 (0.213, 0.253, 0.339, 0.194)。在这一较短的时间内，居住地发生的变化非常小。上述检验统计量的值很大，主要反映了庞大样本规模的影响。为了估计某个地区的居住地变动情况，我们将所有其他地区进行合并，然后对新形成的 2 × 2 表格应用式 10.2。由此可得，关于 $\pi_{+1} - \pi_{1+}$ 的 95% 的置信区间等于 $(0.2131 - 0.2179) \pm 1.96(0.00054)$，即 -0.005 ± 0.001。相似地，关于 $\pi_{+2} - \pi_{2+}$ 的 95% 的置信区间为 -0.007 ± 0.001，$\pi_{+3} - \pi_{3+}$ 的相应区间为 0.009 ± 0.001，$\pi_{+4} - \pi_{4+}$ 的相应区间为 0.003 ± 0.001。尽管统计结果表明所有四个地区都发生了明显变化，但是实质上这些变化非常微小。

10.4　对称性、准对称性以及准独立性

方形列联表的另一种分析策略是利用 Logit 模型或对数线性模型直接分析它的联合分布。边际同质性模型是其中一些模型的特例。

如果一个 $I \times I$ 的联合分布 $\{\pi_{ab}\}$ 对于任意 $a \neq b$ 都存在

$$\pi_{ab} = \pi_{ba}, \qquad (10.17)$$

那么我们就称其满足*对称性*(*symmetry*)。在这种情况下，对所有 a 都有 $\pi_{a+} = \sum_b \pi_{ab} = \sum_b \pi_{ba} = \pi_{+a}$，也即，它也同时满足边际同质性。当 $I = 2$ 时，对称性等价于边际同质性，但当 $I > 2$ 时，在非对称的情况下也可能存在边际同质性。

10.4.1　关于对称性的 Logit 和对数线性模型

当所有 $\pi_{ab} > 0$ 时，对称性可以表示为 Logit 模型和对数线性模型的形式。按照 Logit 的形式，对称性模型可以简单表示为：

$$\text{对于所有 } a < b, \log(\pi_{ab}/\pi_{ba}) = 0。$$

对于对称性表格的期望频数 $\{\mu_{ab} = n\pi_{ab}\}$，其对数线性形式为

$$\log \mu_{ab} = \lambda + \lambda_a + \lambda_b + \lambda_{ab}, \tag{10.18}$$

其中所有 $\lambda_{ab} = \lambda_{ba}$。两个变量具有相同的单变量参数 $\{\lambda_a\}$，因而 $\log \mu_{ab} = \log \mu_{ba}$。模型的可识别性要求设置一定的约束条件。该模型的一种更简单的表述为 $\log \mu_{ab} = \lambda_{ab}$，并且所有 $\lambda_{ab} = \lambda_{ba}$。

对于服从泊松分布或者多项分布的单元格计数 $\{n_{ab}\}$，似然方程为：

$$\text{对所有 } a < b, \hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba}; \text{对所有 } a, \hat{\mu}_{aa} = n_{aa}。$$

列联表的主对角线上存在完全的拟合。满足对称性的解为：

$$\text{对所有 } a \text{ 和 } b, \hat{\mu}_{ab} = \frac{n_{ab} + n_{ba}}{2}。$$

Logit 形式的对称性模型不包括关于 $a < b$ 的 $\binom{I}{2}$ 个二项分布配对 $\{(n_{ab}, n_{ba})\}$ 的参数,因而它的残差自由度等于 $df = I(I-1)/2$。等价地,对数线性形式的对称性模型 $\log \mu_{ab} = \lambda_{ab}(\lambda_{ab} = \lambda_{ba})$ 将 $\{n_{ab}\}$ 视为 I^2 个泊松分布计数,模型在 $a < b$ 时具有 $\binom{I}{2}$ 个 $\{\lambda_{ab}\}$ 参数,再加上 I 个 $\{\lambda_{aa}\}$ 参数,因而相应的残差自由度为 $df = I^2 - [I + I(I-1)/2] = I(I-1)/2$。Bowker(1948)表明,在进行对称性检验时, X^2 可简化为：

$$X^2 = \sum_{a<b}\sum \frac{(n_{ab} - n_{ba})^2}{n_{ab} + n_{ba}}。$$

当 $I = 2$ 时,上式为 McNemar 统计量,即公式 10.4 的平方。模型的标准化皮尔逊残差等于

$$r_{ab} = (n_{ab} - n_{ba})/(n_{ab} + n_{ba})^{1/2}。$$

由于 $r_{ab} = -r_{ba}$，所以对于每对类别而言,只有一个残差是非冗余的。这些残差满足 $\sum_{a<b} r_{ab}^2 = X^2$。

对称性模型非常简单。除了少数特定应用,如描述同一观测员在进行重复测量时所得结果间的一致性外,对称性模型的拟合结果一般都很差。尤其是当两个变量的边际分布差别很大时,对称性模型拟合得很差。

10.4.2 准对称性

通过允许对称性模型式 10.18 中的主效应项取不同值,可以处理边际异质性的情况。这样所得到的对数线性模型被称为*准对称性*(*quasi-symmetry*)模型,表示为

$$\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}, \tag{10.19}$$

其中对于任意 $a < b$，有 $\lambda_{ab} = \lambda_{ba}$ (Caussinus,1966)。对称性模型是准对称性模型中 $\lambda_a^X = \lambda_a^Y$ 时的一个特例,其中 $a = 1,\cdots,I$; 独立性模型则是准对称性模型中所有 $\lambda_{ab} = 0$ 时的一个特例。

准对称性模型的似然方程是：

$$\begin{aligned} \hat{\mu}_{a+} &= n_{a+}, \quad a = 1,\cdots,I, \\ \hat{\mu}_{+b} &= n_{+b}, \quad b = 1,\cdots,I, \\ \text{对 } a \leqslant b, \hat{\mu}_{ab} &+ \hat{\mu}_{ba} = n_{ab} + n_{ba}。 \end{aligned} \tag{10.20}$$

前两组方程中只有一组是必须的:在给定其他两组方程后,另一组是冗余的。模型的残差自由度等于 $df = (I-1)(I-2)/2$。由 10.20 可得,对于 $a = 1,\cdots,I$, $\hat{\mu}_{aa} = n_{aa}$。否则,似然方程不具有直接的解。这些似然方程可以通过迭代法来求解,如 Newton-Raphson 算

法和 IPF 算法(Caussinus,1966)。

准对称性模型具有以下的可积形式:

$$\text{对所有 } a < b, \pi_{ab} = \alpha_a \beta_b \gamma_{ab}, \text{其中 } \gamma_{ab} = \gamma_{ba}, \tag{10.21}$$

其中所有的参数都取正值。在式 10.21 中,当对所有 a 都存在 $\alpha_a = \beta_a$ 时,便得到了对称性模型。这个等式表明,满足准对称性的表格实际上是一个满足独立性的表格与一个满足对称性的表格之间相应单元格两两相乘的结果。对称性关联意味着在主对角线一边的发生比之比与在主对角线另一边的相应发生比之比都相等。事实上,这种模型可以通过以下特性来界定:

$$\text{对所有 } a < b, \frac{\mu_{ab}\mu_{II}}{\mu_{aI}\mu_{Ib}} = \frac{\mu_{ba}\mu_{II}}{\mu_{bI}\mu_{Ia}}, \tag{10.22}$$

或者局部发生比之比满足 $\theta_{ab} = \theta_{ba}$。Goodman(1979a)将此类模型称为*对称关联*(*symmetric association*)模型。

准对称性的含义不如对称性那么明显,但是,它的拟合结果一般比对称性模型好得多,适用范围也更广。准对称性模型中参数的一种解释方法与对象别 Logit 模型有关。对于那些包括可加的对象效应(subject terms)和场合效应(occasion terms)的模型,其中式 10.8 模型是最简单的一个,相应的总体平均表格的联合分布必然满足准对称性(参见:Darroch(1981);本书第 13.2.7 节也给出了这一结果)。考虑将式 10.15 基线类别 Logit 模型扩展为一个对象别模型

$$\log\left[P(Y_{it} = j)/P(Y_{it} = I)\right] = \alpha_{ij} + \beta_j x_t, t = 1,2, \quad j = 1,\cdots,I-1。$$

这个模型对于每个 j 都具有式 10.18 的可加形式。该模型意味着,当约束条件为 $\lambda_I^X = \lambda_I^Y = 0$ 时,对所有对象求平均所得的 $I \times I$ 的总体平均表格满足式 10.19 准对称性模型,其中 $\{\beta_j = \lambda_j^Y - \lambda_j^X\}$。事实上,在消除掉参数 $\{\alpha_{ij}\}$ 的条件最大似然分析中,关于 $\{\hat{\beta}_j\}$ 的条件最大似然估计与关于准对称性模型的普通最大似然拟合具有以下关系:$\{\hat{\beta}_j = \hat{\lambda}_j^Y - \hat{\lambda}_j^X\}$(Conaway,1989)。这一结果可用来解释准对称性模型中的主效应项。

相应结果对多场合情况下的多元准对称性模型(10.33)也成立(如:Agresti(1997);Conaway(1989);Darroch(1981);Tjur(1982);另见第 13.2.7 节)。另外,还有一些有用的其他模型是准对称性模型的特例,其中包括第 10.4.3 和第 10.6.3 节将要介绍的模型。

10.4.3 准独立性

方形表数据通常存在正向相依,表现为主对角线上的单元格计数大于独立性模型的预测值。然而,对于落在主对角线外的配对数据而言,它们之间可能具有非常简单的关联结构。

在给定方形表的行和列的结果不一致的情况下,如果两个变量相互独立,那么称这个方形列联表满足*准独立性*(*quasi-independence*)。准独立性模型的对数线性形式为

$$\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \delta_a I(a = b), \tag{10.23}$$

其中 $I(\cdot)$ 是一个指示函数,

$$I(a = b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}。$$

准独立性模型在独立性模型的基础上对主对角线上的每个单元格加入了一个参数。式 10.23 的前三项设定了独立性,而 $\{\delta_a\}$ 允许 $\{\mu_{aa}\}$ 偏离这一模式并可以取任意的正值。当 $\delta_a > 0$ 时,μ_{aa} 的取值比独立性模型所预测的大。

准独立性模型的似然方程为

$$\hat{\mu}_{a+} = n_{a+}, \quad \hat{\mu}_{+a} = n_{+a}, \quad \hat{\mu}_{aa} = n_{aa}, \quad a = 1, \cdots, I。$$

在准独立性模型中,主对角线上的单元格存在着完全拟合,但其他的单元格满足独立性。这一模型意味着,由任意非主对角线上的单元格所组成的 2×2 表格的发生比之比等于 1.0。准独立性模型可以通过 Newton-Raphson 算法或 IPF 算法来拟合。它比独立性模型多包括 I 个参数,因而其残差自由度为 $\text{df} = (I-1)^2 - I$。准独立性模型同样也适用于 $I \geq 3$ 的表格。

准独立性模型是式 10.21 准对称性模型在 $a \neq b$ 时所有 $\{\gamma_{ab}\}$ 都相等的一个特例。Caussinus(1966:146)表明,当 $I = 3$ 时,两个模型是等价的。

10.4.4 例子:再析迁移数据

现在,我们回到表 10.6 有关迁移模式的数据。不出所料,独立性模型对这些数据拟合得很差,它的 $G^2 = 125\,923$, $X^2 = 146\,929$ (X^2 的最大可能取值为 $3n = 167\,943$; 参见习题 3.33)。对称性模型的拟合结果也不能令人满意。例如,表中有 124 人从东北部迁移到西部,而仅有 63 人进行了反向迁移。检验对称性模型的偏离度为 $G^2 = 243.6$ (df = 6)。

准独立性模型指的是,对于那些发生了迁移的人们来说,他们在 1985 年的居住地独立于其在 1980 年的居住地。表 10.7 给出了该模型的拟合值,它的 $G^2 = 69.5$ (df = 5)。该模型的拟合结果比独立性模型好得多,这主要是由于准独立性模型对主对角线上的计数强行进行了完全拟合,而大多数观测值落在了主对角线上。不过,对非主对角线上的计数,准独立性模型明显拟合得不充分。与该模型的预测相比,更多的人从东北部迁移到了南部,而较少的人从西部迁移到了南部。

表 10.7 关于表 10.6 数据的模型拟合结果

1980 年的居住地	1985 年的居住地[a]				小计
	东北部	中西部	南部	西部	
东北部	11 607	100	366	124	12 197
		(126.6)[1]	(312.9)	(150.5)	
		(95.8)[2]	(370.4)	(123.8)	
中西部	87	13 677	515	302	14 581
	(117.4)		(531.1)	(255.5)	
	(91.2)		(501.7)	(311.1)	
南部	172	255	17 189	270	18 486
	(133.2)	(243.8)		(290.0)	
	(167.6)	(238.3)		(261.1)	
西部	63	176	286	10 192	10 717
	(71.4)	(130.6)	(323.0)		
	(63.2)	(166.9)	(294.9)		
小计	11 929	14 178	18 986	10 888	55 981

a 1 准独立性模型的拟合值;2 准对称性模型的拟合值;两个模型在主对角线上存在完全拟合。

准对称性模型具有 $G^2 = 3.0$ (df = 3)。表 10.7 也给出了准对称性模型的拟合结果,它比准独立性模型的拟合结果好得多。单元格概率之间缺乏对称性反映了数据存在微小的边际异质性。可以利用准对称性模型的参数估计来描述相应的对象别效应, $\{\hat{\lambda}_1^Y - \hat{\lambda}_1^X = -0.672, \quad \hat{\lambda}_2^Y - \hat{\lambda}_2^X = -0.623, \quad \hat{\lambda}_3^Y - \hat{\lambda}_3^X = 0.122\}$。例如,给定一个对象,估计其在

1985 年居住在南部而不是西部的发生比是1980 的相应发生比的 exp(0.122) = 1.13 倍。在第 12 章我们将看到,这些对象别效应往往比相应边际模型中的效应更强,尤其是当表格中存在强关联时。

与配对样本有关的一项应用是关于职业流动的研究,其中每个观测值都是由父母的职业和孩子的职业形成的配对(Goodman,1979b;Hout et al.,1987)。

10.4.5 边际同质性与准对称性

边际同质性本身不等价于一个对数线性模型。但是,准对称性模型在分析边际同质性时非常有用。Caussinus(1966)证明,对称性等价于准对称性和边际同质性同时成立。我们在前文已经看到,对称性既意味着准对称性,也意味着边际同质性。现在,我们给出 Caussinus 的论断的逆命题,准对称性和边际同质性同时成立意味着对称性。

由式 10.21 可得,如果准对称性成立,那么 $\pi_{ab} = \alpha_a\beta_b\gamma_{ab}$,其中对所有 $a < b$, $\gamma_{ab} = \gamma_{ba} > 0$。等价地,

$$\pi_{ab} = \rho_a\delta_{ab},$$

其中 $\rho_a = \alpha_a/\beta_a$, $\delta_{ab} = \beta_a\beta_b\gamma_{ab}$ 且对所有 $a < b$ 满足 $\delta_{ab} = \delta_{ba} > 0$。如果边际同质性也成立,那么

$$\pi_{j+} = \rho_j\sum_b\delta_{jb} = \sum_a\rho_a\delta_{aj} = \pi_{+j},$$

或者

$$\rho_j = \sum_a\rho_a\delta_{aj}\Big/\sum_b\delta_{jb} = \sum_a\rho_a\delta_{aj}\Big/\sum_b\delta_{bj}, \quad j = 1,\cdots,I。$$

因此,每个 ρ_j 都是关于 $\{\rho_a\}$ 的加权平均,权数为 $\{\delta_{aj}/\sum_b\delta_{bj} > 0, a = 1,\cdots,I\}$。任何一组满足以上条件的 $\{\rho_a\}$ 必然都相等,否则,一定会存在一个 ρ_j 不大于任意的 ρ_a 但至少小于其中的一个,因而它不可能等于对所有 ρ_a 的正向加权平均。但是由于 $\{\rho_a\}$ 是恒等的, $\pi_{ab} = \rho_a\delta_{ab} = \rho_b\delta_{ab} = \pi_{ba}$,所以对称性成立。因此,一个既满足准对称性又满足边际同质性的表格也同时满足对称性。由于逆命题成立,所以

$$\text{准对称性} + \text{边际同质性} = \text{对称性}。\tag{10.24}$$

由此推出,当准对称性(QS)成立时,边际同质性(MH)等价于对称性(S),也即在 QS 模型中具有 $\{\lambda_a^X = \lambda_a^Y, a = 1,\cdots,I\}$。因此,以准对称性为条件,相应的边际同质性检验等价于关于对称性的检验。通过比较对称性模型和准对称性模型的拟合优度统计量,我们可以对边际同质性进行检验:

$$G^2(\mathrm{S}\mid\mathrm{QS}) = G^2(\mathrm{S}) - G^2(\mathrm{QS}),\tag{10.25}$$

检验的自由度为 df = $I - 1$。这是除第 10.3.3 节介绍的边际模型外另一种进行边际同质性检验的方法。

在表 10.6 有关 1980—1985 年迁移状况的数据中, $G^2(\mathrm{S}) = 243.6$, $G^2(\mathrm{QS}) = 3.0$。二者之差 $G^2(\mathrm{S}|\mathrm{QS}) = 240.6$(df = 3)表明,该数据具有很强的边际异质性。这一结果与第 10.3.4 节所引述的基于式 10.15 模型的检验结果相似,其中似然比检验的 $G^2 = 240.8$,沃尔德检验的 $W = 236.5$(自由度均为 df = 3)。

10.4.6 定序准对称性模型

到目前为止,我们所介绍的关于方形表的对数线性模型都将变量视为定类测度的。当变量的类别间存在排序时,更为简约的模型具有特别意义。令 $u_1 \leqslant \cdots \leqslant u_I$ 表示关于

行和列的类别赋值。*定序准对称性模型*（*ordinal quasi-symmetry model*）可表示为

$$\log \mu_{ab} = \lambda + \lambda_a + \lambda_b + \beta u_b + \lambda_{ab}, \quad (10.26)$$

其中对所有 $a < b$ 存在 $\lambda_{ab} = \lambda_{ba}$。它是式 10.19 准对称性模型满足以下线性趋势

$$\lambda_b^Y - \lambda_b^X = \beta u_b$$

时的一个特例。对称性模型则是上式在 $\beta = 0$ 时的一个特例。

定序准对称性模型的 Logit 形式为：

$$\text{对 } a \leqslant b \qquad \log(\pi_{ab}/\pi_{ba}) = \beta(u_b - u_a)。\quad (10.27)$$

这是线性 Logit 模型 Logit$(\pi) = \alpha + \beta x$ 的一个特例，其中 $\alpha = 0$，$x = u_b - u_a$，π 表示给定结果为 (a,b) 或 (b,a) 之一，但最终结果为 (a,b) 的条件概率。$|\beta|$ 的值越大，π_{ab} 和 π_{ba} 之间的差异就越大，进而两组边际分布之间的区别也越大。

定序准对称性模型的似然方程是

$$\sum_a u_a \hat{\mu}_{a+} = \sum_a u_a n_{a+}, \qquad \sum_b u_b \hat{\mu}_{+b} = \sum_b u_b n_{+b},$$

$$(\text{对于 } a < b) \qquad \hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba}。$$

该模型所拟合的边际计数不一定要等于观察到的边际计数。但是，将前两个方程除以 n 表明，拟合值和观察值之间具有相同的均值。

当 $\beta \neq 0$ 时，定序准对称性模型意味着边际分布存在随机的排序。当 $\beta > 0$（或 $\beta < 0$）时，列（或行）分布中的取值具有较大的均值。与定序边际模型（第 10.3.1 节）一样，这个模型通过 df = 1 个自由度来集中考察边际效应。关于边际同质性（$H_0:\beta = 0$）的检验可以利用等价条件：

$$\text{定序准对称性} + \text{边际同质性} = \text{对称性},$$

即比较对称性模型和定序准对称性模型偏离度之差的似然比检验统计量。

大家可以利用拟合式 10.27Logit 模型的有关软件来拟合定序准对称性模型：将 (n_{ab}, n_{ba}) 视为 $n_{ab} + n_{ba}$ 次试验的二项分布结果，拟合一个不包括截距项、预测变量为 $x = u_b - u_a$ 的 Logit 模型。此外，也可以通过迭代法拟合相应的式 10.26 对数线性模型。

10.4.7　例子：再论婚前性行为和婚外性行为

对于表 10.5 有关婚前和婚外性行为态度的数据，粗略地看对称性模型是不充分的（$G^2 = 402.2$，df = 6）。相对而言，准对称性模型对数据拟合得很好（$G^2 = 1.4$，df = 3）。更简单的定序准对称性模型也拟合得很好：利用赋值 {1,2,3,4}，$G^2 = 2.1$（df = 5）。

定序准对称模型的最大似然估计 $\hat{\beta} = -2.86$。根据式 10.27 可知，所估计的对婚前性行为的态度比对婚外性行为更正面 x 个类别的概率是相反概率的 $\exp(2.86x)$ 倍。例如，认为婚前性行为"几乎总是错的"而婚外性行为"总是错的"的概率，是认为婚前性行为"总是错的"而婚外性行为"几乎总是错的"的概率的大约 $\exp(2.86) = 17.4$ 倍。

10.4.8　方形表格的其他定序模型

在定序划分的情况下，当对称性模型不成立时，对所有 $a < b$，经常存在 $\pi_{ab} > \pi_{ba}$ 或者 $\pi_{ab} < \pi_{ba}$。就这一数据特点将对称性加以扩展，得到 Logit 模型：

$$\text{对 } a < b \qquad \log(\pi_{ab}/\pi_{ba}) = \tau。\quad (10.28)$$

它意味着，对所有 $a < b$，

$$P(Y_{i1} = a, Y_{i2} = b \mid Y_{i1} < Y_{i2}) = P(Y_{i1} = b, Y_{i2} = a \mid Y_{i1} > Y_{i2})。$$

即，主对角线以上的单元格的概率模式与主对角线以下的相同。这种特性被称为*条件对*

称性(*conditional symmetry*)(McCullagh,1978)。习题 10.35 给出了相应的对数线性模型及其拟合。对称性模型是其中 $\tau = 0$ 时的一个特例。

另一个模型是对准独立性模型的扩展。令 $\{u_a\}$ 表示排序的赋值。模型

$$\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \beta u_a u_b + \delta_a I(a = b) \tag{10.29}$$

允许主对角线以外的元素具有双线性关联(参见式9.6)。这个模型是准对称性模型的一个特例,而准独立性模型是上述模型中 $\beta = 0$ 时的一个特例。在使用等距赋值的情况下,该模型意味着,给定结果变量的不同取值,存在一致性局部关联。Goodman(1979a)称其为*准一致性关联*(*quasi-uniform association*)模型。

对于表 10.5 中有关婚前和婚外性行为态度的数据,条件对称性模型具有 $\hat{\tau} = -4.130$ (SE = 0.451)。模型所估计的更反对婚外性行为的概率是更反对婚前性行为的相应概率的 exp(4.13) = 62.2 倍。准一致性关联模型具有 $\hat{\beta} = 0.632$ (SE = 0.106),因而,除主对角线上的单元格外,局部发生比之比的估计值等于 exp(0.632) = 1.88。

10.5 不同评定者之间评定结果的一致性

接下来,我们讨论关于配对模型的一种应用——分析两个评定者之间评定结果的一致性。以表 10.8 为例,该表给出了两位病理学家,记为 *A* 和 *B*,分别对 118 张幻灯片做出的关于是否出现子宫颈癌及其严重程度的评定结果。评定尺度具有以下排序类别:(1)阴性,(2)异常鳞状增生,(3)原位癌,(4)鳞状或浸润性癌。

表 10.8 关于子宫颈癌的诊断

病理学家 A	病理学家 B[a]				小计
	1	2	3	4	
1	22	2	2	0	26
	(8.5)	(−0.5)	(−5.9)	(−1.8)	
2	5	7	14	0	26
	(−0.5)	(3.2)	(−0.5)	(−1.8)	
3	0	2	36	0	38
	(−4.1)	(−1.2)	(5.5)	(−2.3)	
4	0	1	17	10	28
	(−3.3)	(−1.3)	(0.3)	(5.9)	
小计	27	12	69	10	118

a 括号中的数值为独立性模型的标准化皮尔逊残差。

来源:N. S. Holmquist, C. A. McMahon, and O. D. Williams, *Arch. Pathol.* 84:334-345(1967);经美国医学会(the American Medical Association)授权重印。另见:Landis and Koch(1977)。

10.5.1 一致性:对独立性的偏离

令 π_{ab} 表示评定者 *A* 将一张幻灯片评定为类别 *a* 而评定者 *B* 将其评定为类别 *b* 的概率,那么,π_{aa} 是他们都选择类别 *a* 的概率,并且 $\sum_a \pi_{aa}$ 表示他们的评定结果一致的总概率。当 $\sum_a \pi_{aa} = 1$ 时,两个评定者的评定全部一致。

在使用主观性评定尺度的情况下,评定者之间很少能达成完全一致。这时,分析的重点是描述他们之间一致性的强度,并找出出现不一致情况的模式。*一致性*(*agreement*)

和关联（*association*）反映的是联合分布的不同方面。强一致性必然要求具有强关联，但是强关联的存在并不意味着强一致性。如果评定者 A 的评定结果比评定者 B 系统性地高出一个类别，这时尽管二者间存在强关联，但是一致性却很差。

通过比较观察值 $\{n_{ab}\}$ 与独立性模型的预测值 $\{n_{a+}n_{+b}/n\}$，可以用来评估一致性。独立性模型是一个基线模型，表明如果在评定结果之间不存在关联时所预期的一致性。正常情况下，即便评定结果间存在较弱的一致性，独立性模型都拟合得很差，但是它的单元格标准化残差（第 3.3.1 节）能够反映一致性与不一致性的发生模式。理想状态下，主对角线上的标准化残差取很大的正值，而主对角线以外的标准化残差取很大的负值。然而，残差的大小会受到样本规模 n 的影响，随着 n 的增大，残差的值也会变大。

独立性模型对表 10.8 的数据拟合得很差（$G^2 = 118.0$，df = 9）。该表在括号中给出了相应的标准化皮尔逊残差。在主对角线上具有很大的正残差表明，评定结果间每个类别的一致性要比纯随机的情况下所预测的大，尤其是第一个类别。主对角线以外的标准化残差基本都是负的。不一致发生的情况比独立性模型所预测的低，尽管在相近类别之间这一特征相对较弱。最普遍的不一致性出现在，当评定者 B 选择类别 3 时，评定者 A 却选择类别 2 或 4。

10.5.2 通过准独立性模型来分析一致性

在独立性模型的基础上加入与一致性有关的项，便得到更复杂的模型。准独立性模型式 10.23 就是一个有用的扩展，它在独立性模型中增加了主对角线参数 $\{\delta_a\}$。对于表 10.8 的数据，准独立性模型具有 $G^2 = 13.2$（df = 5）。它的拟合结果比独立性模型好得多，但仍然存在一定的拟合不足。表 10.9 给出了准独立性模型的拟合结果。

表 10.9 关于表 10.8 子宫颈癌诊断数据的拟合值

病理学家 A	病理学家 B[a]			
	1	2	3	4
1	22	2	2	0
	(22)[1]	(0.7)	(3.3)	(0.0)
	(22)[2]	(2.4)	(1.6)	(0.0)
2	5	7	14	0
	(2.4)	(7)	(16.6)	(0.0)
	(4.6)	(7)	(14.4)	(0.0)
3	0		36	0
	(0.8)	(1.2)	(36)	(0)
	(0.4)	(1.6)	(36)	(0.0)
4	0	1	17	10
	(1.9)	(3.0)	(13.1)	(10)
	(0.0)	(1.0)	(17.0)	(10)

a 1 准独立性模型；2 准对称性模型。

以两个研究对象为例，假如每个评定者分别将其中一个划分在类别 a，另一个划分在类别 b。两个评定者的评定结果一致而不是相矛盾的发生比等于

$$\tau_{ab} = \frac{\pi_{aa}\pi_{bb}}{\pi_{ab}\pi_{ba}} = \frac{\mu_{aa}\mu_{bb}}{\mu_{ab}\mu_{ba}}。\tag{10.30}$$

随着 τ_{ab} 的上升，评定者之间也就更可能对这两个类别的划分达成一致。在准独立性模型下，

$$\tau_{ab} = \exp(\delta_a + \delta_b)。$$

$\{\delta_a\}$ 越大,表示一致性越强。例如,就表 10.8 的数据而言,$\hat{\delta}_2 = 0.6$,$\hat{\delta}_3 = 1.9$,且 $\hat{\tau}_{23} = 12.3$。其他几对类别间的一致性也还不错。

10.5.3 准对称性与一致性模型

对于表 10.8,准独立性模型的拟合结果显示出一定的拟合不足。给定病理学家评定结果不一致的情况下,在这些不一致的结果之间仍存在某种关联。这一现象对于评定一致性的表格来说是很常见的。由于准对称性模型式 10.19 允许关联的存在,它通常会拟合得更好。就表 10.8 来说,准对称性模型具有 $G^2 = 1.0$ (df=2)。表 10.9 也给出了该模型的拟合结果。在不少情况下,表格可能出现包括多个空单元格的问题。当任何一对类别(比如表 10.8 中的类别 1 和 4)存在 $n_{ab} + n_{ba} = 0$ 时,准对称性模型关于这些单元格的最大似然拟合值也必然等于零,因为它的似然方程要求 $\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba}$。在拟合过程中应当去除这些单元格,以得出正确的残差自由度。

在准对称性模型下,$\hat{\tau}_{ab} = \exp(\hat{\lambda}_{aa} + \hat{\lambda}_{bb} - \hat{\lambda}_{ab} - \hat{\lambda}_{ba})$,其中 $\hat{\lambda}_{ab} = \hat{\lambda}_{ba}$。例如,对于表 10.8 中的类别 2 和 3,$\hat{\tau}_{23} = 10.7$。

对数线性模型能够直接估计关于一致性的关联项。准对称性模型还给出了评定结果的边际分布是否相似的信息。更简单的对称性模型强行要求满足边际同质性,它对表 10.8 的拟合很差($G^2 = 39.2$, df=5)。检验统计量 $G^2(S|QS) = 39.2 - 1.0 = 38.2$(df=3)给出了存在边际异质性的有力证据。在表 10.8 中,除第一个类别以外,其他类别的边际比例相差很大。由此可见,边际异质性是评定结果间总体一致性不强的重要原因。

关于一致性的模型可以将类别间的排序特征考虑进来。以评定结果相矛盾的情况为条件,通常会存在这样一种趋势:当一个评定者所给的评分高(低)时,另一个评定者的评分也较高(低)(参见习题 10.41)。

10.5.4 测量一致性的卡帕指标

分析一致性的另一种方法是用单个指标来反映一致性的总体情况。当变量为定类变量时,最常用的测量指标是 *Cohen* 的卡帕(*Cohen's kappa*)(Cohen,1960)。它将观察数据中出现一致性的概率 $\sum_a \pi_{aa}$ 与独立性模型的期望概率 $\sum_a \pi_{a+}\pi_{+a}$ 相比较,

$$\kappa = \frac{\sum_a \pi_{aa} - \sum_a \pi_{a+}\pi_{+a}}{1 - \sum_a \pi_{a+}\pi_{+a}}。$$

其中,分母中的 1 代表完全一致性,即分子中 $\sum_a \pi_{aa}$ 的最大可能取值。当一致性仅等于独立性模型的期望值时,卡帕等于 0;当出现完全一致性时,卡帕等于 1.0。给定评定结果的边际分布,一致性的程度越强,κ 的值也就越高。当一致性比纯粹偶然的情况还要差时,卡帕取负值,但这种情况几乎不会发生。

对于多项分布样本,样本值 $\hat{\kappa}$ 服从大样本正态分布,其渐近方差的估计公式(Fleiss et al.,1969)为

$$\hat{\sigma}^2(\hat{\kappa}) = \frac{1}{n}\left\{\frac{P_o(1-P_o)}{(1-P_e)^2} + \frac{2(1-P_o)\left[2P_oP_e - \sum_a p_{aa}(p_{a+}+p_{+a})\right]}{(1-P_e)^3} + \frac{(1-P_o)^2\left[\sum_a\sum_b p_{ab}(p_{b+}+p_{+a})^2 - 4P_e^2\right]}{(1-P_e)^4}\right\},$$

其中 $P_o = \sum_a p_{aa}$，$P_e = \sum_a p_{a+}p_{+a}$。一致性的水平还不如纯随机情况的可能性很小。因此，与假设检验 $H_0:\kappa = 0$ 相比，通过对 κ 的区间估计来描述一致性的强度更有意义。

对于表 10.8，$P_o = 0.636$，$P_e = 0.281$。样本的卡帕等于 $(0.636 - 0.281)/(1 - 0.281) = 0.493$，也即，所观察到的一致性与独立性模型的期望值之差大约是二者最大可能差异值的 50%。相应的标准误估计值为 0.057，所以 κ 大约落在 0.4 至 0.6 之间，表现出中等强度的一致性。

10.5.5 加权卡帕：对不一致性加以量化

卡帕将变量的划分尺度视为定类的。当类别间存在排序时，不一致性的严重程度取决于评定结果之间的差别。即使在定类划分的情况下，评定结果不一致的严重程度仍可能不尽相同。*加权卡帕*（*weighted kappa*）（Spitzer et al.，1967）利用满足 $0 \leqslant w_{ab} \leqslant 1$ 的权数 $\{w_{ab}\}$ 来描述一致性的接近程度，其中所有 $w_{aa} = 1$ 且 $w_{ab} = w_{ba}$。一种可选的权数赋值为 $\{w_{ab} = 1 - |a-b|/(I-1)\}$，这里，越接近主对角线的单元格一致性越强。Fleiss 和 Cohen（1973）建议使用 $\{w_{ab} = 1 - (a-b)^2/(I-1)^2\}$。加权后的一致性为 $\sum_a \sum_b w_{ab}\pi_{ab}$，并且加权卡帕等于

$$\kappa_w = \frac{\sum_a \sum_b w_{ab}\pi_{ab} - \sum_a \sum_b w_{ab}\pi_{a+}\pi_{+b}}{1 - \sum_a \sum_b w_{ab}\pi_{a+}\pi_{+b}}。$$

关于卡帕和加权卡帕的应用价值存在很多争议，部分原因在于它们的取值严重依赖于边际分布的情况。在各种类别的案例所占比例不同的情况下，同样的诊断评定过程得出的卡帕值会存在很大差异（习题 10.40）。在使用单个指标来综合描述整个列联表时，可能会损失非常重要的信息。因此，通过构建模型来考察一致性与不一致性的具体结构，比仅依赖于一个综合性指标更有意义。

10.5.6 扩展到多个评定者的情况

当存在多个评定者时，普通的对数线性模型一般没有什么价值，因为它对两个评定者之间的一致性和关联的描述是以其他评定者的评定结果为条件的。这时，分析无条件的边际关联更有意义。因而，在具有 R 个评定者时，同时对每对评定者的一致性和关联结构进行模型分析，总共需要考察 $\binom{R}{2}$ 对二维边际分布（Becker and Agresti，1992）。

在这种情况下，可以考虑使用其他方法。例如，一般化的卡帕指标可以用来测量成对的一致性或多元一致性（Fleiss，1981，sec. 13.2；Landis and Koch，1977）。或者，也可以通过假定评定者评定结果一致与不一致的对象分别属于不同潜类（latent class），拟合相应的混合模型。有关分析，详见第 13.1.2 节。

10.6 关于成对选择的 BRADLEY-TERRY 模型

有时，分类结果变量来自于成对的评估比较。一个常见的例子是体育比赛，一个队或运动员的比赛结果分为（赢，输）。另一个例子是关于产品品牌的成对比较，比如某种酒类的两个品牌。当一位品酒师品评 I 种白索维农酒时，他很难给出一个完整的排序，尤其是在 I 很大的情况下。但是，对于其中的任意两种，品酒师在同一场合进行品尝后可以

给出他的偏好。我们可以基于这种两两比较,得出关于所有酒的总排序。在本节中,我们介绍一个与此有关的模型。

10.6.1 Bradley-Terry 模型

Bradley 和 Terry(1952)提出了一个分析成对比较数据的 Logit 模型。令 $\prod_{ab}$ 表示对 a 的偏好超过对 b 的偏好的概率。假定对于任意一对选择,都有 $\prod_{ab}+\prod_{ba}=1$,也即,不存在偏好相同的情况。Bradley-Terry 模型可表示为

$$\log\frac{\prod_{ab}}{\prod_{ba}}=\beta_a-\beta_b。\tag{10.31}$$

或者

$$\prod_{ab}=\exp(\beta_a)/[\exp(\beta_a)+\exp(\beta_b)]。$$

因此,当 $\beta_a=\beta_b$ 时,$\prod_{ab}=\frac{1}{2}$;当 $\beta_a>\beta_b$ 时,$\prod_{ab}>\frac{1}{2}$。

模型的可识别性条件可以是 $\beta_I=0$ 或者 $\sum_a\exp(\hat{\beta}_a)=1$。由于该模型利用 $(I-1)$ 个参数描述了 $\binom{I}{2}$ 个概率($a<b$ 时的 $\{\prod_{ab}\}$),它的残差自由度为 $\mathrm{df}=\binom{I}{2}-(I-1)$。

当 $a<b$ 时,令 N_{ab} 表示进行比较的样本数量,其中选择 a 的次数为 n_{ab},选择 b 的次数为 $n_{ba}=N_{ab}-n_{ab}$。这些结果可以通过一个主对角线上的单元格为零的方形列联表来表示。当 N_{ab} 次比较是独立的,并且每次的概率为 $\prod_{ab}$ 时,n_{ab} 服从二项分布 $\mathrm{bin}(N_{ab},\prod_{ab})$。如果有关不同配对的比较也是独立的,那么可以利用拟合普通的 Logit 模型的方法来拟合此模型。

10.6.2 棒球比赛中的主场优势

表 10.10 给出了 1987 赛季美国棒球联盟东区的七支球队的战况。例如,在波士顿与纽约的比赛中,波士顿赢了 7 场,纽约赢了 6 场。表 10.10 统计的是整个常规赛的所有比赛的情况。我们将此数据视为代表按照 1987 年状况所构建的球队长期表现的假想分布的一个样本估计。

我们对 $\binom{7}{2}=21$ 个独立的二项分布样本拟合 Bradley-Terry 模型,在拟合过程中将其视为应用适当模型矩阵并且不包括截距项的 Logit 模型(有关 SAS 程序,参见附表 A. 19)。该模型对数据的拟合很充分($G^2=15.7$, df=15)。表 10.10 给出了相应的拟合值 $\{\hat{\mu}_{ab}\}$。表 10.11 显示了每个球队在比赛中获胜的样本比例以及模型的估计值 $\{\hat{\beta}_a\}$(设定 $\hat{\beta}_7=0$)和 $\{\exp(\hat{\beta}_a)\}$(设定 $\sum_a\exp(\hat{\beta}_a)=1$)。当波士顿与纽约比赛时,所估计的波士顿获胜的概率为

$$\hat{\prod}_{54}=\exp(\hat{\beta}_5)/\left[\exp(\hat{\beta}_5)+\exp(\hat{\beta}_4)\right]=0.46。$$

每个 $\hat{\beta}_a$ 以及 $\hat{\beta}_a-\hat{\beta}_b$ 的标准误大约等于 0.3,所以在排名前五的球队之间并不存在明显差异。

表 10.10 美国棒球联盟球队在 1987 赛季的战况

获胜球队	落败球队[a]						
	Milwaukee	Detroit	Toronto	New York	Boston	Cleveland	Baltimore
Milwaukee	—	7(7.0)	9(7.4)	7(7.6)	7(8.0)	9(9.2)	11(10.8)
Detroit	6(6.0)	—	7(7.0)	5(7.1)	11(7.6)	9(8.8)	9(10.5)
Toronto	4(5.6)	6(6.0)	—	7(6.7)	7(7.1)	8(8.4)	12(10.2)
New York	6(5.4)	8(5.9)	6(6.3)	—	6(7.0)	7(8.3)	10(10.1)
Boston	6(5.0)	2(5.4)	6(5.9)	7(6.0)	—	7(7.9)	12(9.8)
Cleveland	4(3.8)	4(4.2)	5(4.6)	6(4.7)	6(5.1)	—	6(8.6)
Baltimore	2(2.2)	4(2.5)	1(2.8)	3(2.9)	1(3.2)	7(4.4)	—

a 括号中的数值是 Bradley-Terry 模型的拟合值。

来源:*American League Red Book*,1988(St. Louis,Mo:Sporting News Publishing Co.)。

表 10.11 对棒球比赛数据拟合 Bradley-Terry 模型的结果

球队	获胜百分比	$\hat{\beta}_i$ (10.31)	$\exp(\hat{\beta}_i)$ (10.31)	$\exp(\hat{\beta}_i)$ (10.32)
Milwaukee	64.1	1.58	0.218	0.220
Detroit	60.2	1.44	0.189	0.190
Toronto	56.4	1.29	0.164	0.164
New York	55.1	1.25	0.158	0.157
Boston	51.3	1.11	0.136	0.137
Cleveland	39.7	0.68	0.089	0.088
Baltimore	23.1	0.00	0.045	0.044

这个模型本身没有考虑哪个球队是主场作战。但对大多数体育比赛来说,都存在一定的主场优势:当一个球队在主场比赛时,它获胜的可能性更大。表 10.12 按照(主队,客队)对 1987 赛季的比赛结果进行了划分。例如,当波士顿作为主队时,它在对阵纽约时胜了 4 场,输了 2 场;当纽约作为主队时,它胜了波士顿 4 场,输了 3 场。现在,对于所有的 $a \neq b$,令 $\prod_{ab}^{*}$ 表示球队 a 战胜球队 b 的概率,其中 a 为主队。考虑 Logit 模型

$$\log \frac{\prod_{ab}^{*}}{1-\prod_{ab}^{*}} = \alpha + (\beta_a - \beta_b)。 \tag{10.32}$$

表 10.12 1987 赛季分主客场的比赛结果

主队	客队						
	Milwaukee	Detroit	Toronto	New York	Boston	Cleveland	Baltimore
Milwaukee	—	4-3	4-2	4-3	6-1	4-2	6-0
Detroit	3-3	—	4-2	4-3	6-0	6-1	4-3
Toronto	2-5	4-3	—	2-4	4-3	4-2	6-0
New York	3-3	5-1	2-5	—	4-3	4-2	6-1
Boston	5-1	2-5	3-3	4-2	—	5-2	6-0
Cleveland	2-5	3-3	3-4	4-3	4-2	—	2-4
Baltimore	2-5	1-5	1-6	2-4	1-6	3-4	—

来源:*American League Red Book*,1988(St. Louis,Mo:Sporting News Publishing Co.)。

当 $\alpha > 0$ 时,表明存在主场优势。两个水平相当的球队中,主队获胜的概率为 $\exp(\alpha)/[1 +$

$\exp(\alpha)$]。

对于表 10.12 的数据，式 10.32 模型利用 7 个参数描述了 42 个二项分布，它具有 $G^2 = 38.6$ (df = 35)。表 10.11 给出了相应的 $\{\exp(\hat{\beta}_a)\}$，结果与前面的模型很相似。关于主场的参数估计为 $\hat{\alpha} = 0.302$。在两个水平相当的球队比赛时，所估计的主队获胜的概率为 0.575。当波士顿队与纽约队比赛时，如果比赛在波士顿举行，那么波士顿队获胜的概率估计为 0.54；如果比赛在纽约举行，那么波士顿队获胜的概率估计为 0.39。

当存在某种*排序效应*(*order effect*)时，式 10.32 模型是对 Bradley-Terry 模型的一种重要扩展。例如，在成对的品评测定中，先品尝的产品可能会具有微小的优势。

10.6.3 Bradley-Terry 模型与准对称性模型

Fienberg 和 Larntz(1976)表明，Bradley-Terry 模型是对式 10.19 准对称性模型的一种 Logit 表述。在满足准对称性的情况下，给定一个观测值落在单元格 (a,b) 或者 (b,a) 中，那么最终结果为单元格 (a,b) 的条件概率的 Logit 等于

$$\log\frac{\mu_{ab}}{\mu_{ba}} = (\lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}^{XY}) - (\lambda + \lambda_b^X + \lambda_a^Y + \lambda_{ba}^{XY})$$

$$= (\lambda_a^X - \lambda_a^Y) - (\lambda_b^X - \lambda_b^Y) = \beta_a - \beta_b,$$

其中 $\beta_a = \lambda_a^X - \lambda_a^Y$。通过准对称性模型的估计值 $\{\hat{\lambda}_a^X\}$ 和 $\{\hat{\lambda}_a^Y\}$，就可以得出 Bradley-Terry 模型的参数估计 $\{\hat{\beta}_a\}$。

10.6.4 对平分情况以及定序评定的扩展

Bradley-Terry 模型可以扩展到进行定序比较的情况，比如在评价两个产品时，使用评定尺度(好得多，稍好一些，相同，稍差，差得多)。当评定尺度包括 I 个类别时，可以利用累积 Logit 模型。令 Y_{ab} 表示在将 a 和 b 进行比较的结果，相应模型为

$$\text{Logit}\left[P(Y_{ab} \leqslant j)\right] = \alpha_j + (\beta_a - \beta_b)。$$

由于 $P(Y_{ab} \leqslant j) = P(Y_{ba} > I - j) = 1 - P(Y_{ba} \leqslant I - j)$，可以推出 $\text{logit}[P(Y_{ab} \leqslant j)] = -\text{Logit}[P(Y_{ba} \leqslant I - j)]$。因此，一定存在 $\alpha_j = -\alpha_{I-j}$。

最常见的定序评定尺度是(胜，平，负)。这时，$\alpha_1 = -\alpha_2$。

10.7 匹配集数据的边际模型和准对称性模型*

分析配对数据的方法可以扩展到匹配集(matched sets)数据的情况。这里，我们主要介绍应用对数线性模型的方法；在第 11 章和第 12 章中，我们将介绍边际模型和条件 Logit 模型方法的相应扩展。

10.7.1 边际同质性、完全对称性以及准对称性

令 $(Y_1, Y_2, \cdots, Y_T)$ 表示每个匹配集中的 T 个结果变量。当结果变量包括 I 个类别时，一个具有 I^T 个单元格的列联表列出了所有的可能结果。令 $\mathbf{i} = (i_1, \cdots, i_T)$ 表示具有 $Y_t = i_t$ 的单元格，$t = 1, \cdots, T$。令 $\pi_{\mathbf{i}} = P(Y_t = i_t, t = 1, \cdots, T)$，并令 $\mu_{\mathbf{i}} = n\pi_{\mathbf{i}}$，则有

$$P(Y_t = j) = \pi_{+\cdots+j+\cdots+},$$

其中下标 j 位于第 t 个变量处，并且 $\{P(Y_t = j), j = 1, \cdots, I\}$ 是 Y_t 的边际分布。

如果对 $j = 1,\cdots,I$，存在

$$P(Y_1 = j) = P(Y_2 = j) = \cdots = P(Y_T = j),$$

那么这个 T 维表格满足*边际同质性*（*marginal homogeneity*）。如果对 $\mathbf{i} = (i_1,\cdots,i_T)$ 的任意排列 $\mathbf{j} = (j_1,\cdots,j_T)$，存在

$$\pi_{\mathbf{i}} = \pi_{\mathbf{j}},$$

那么这个 T 维表格满足*完全对称性*（*complete symmetry*）。完全对称性本身意味着边际同质性，但是除了当 $T = I = 2$ 时，其逆命题不成立。

完全对称性可以通过对数线性模型的形式来表示。它的一种表达形式为

$$\log \mu_{\mathbf{i}} = \lambda_{ab\cdots m},$$

其中 a 是 $(i_1,\cdots,i_T)$ 中的最小值，b 是次小值，……，m 是其中的最大值。例如，在三维表格中，$\log \mu_{122} = \log \mu_{212} = \log \mu_{221} = \lambda_{122}$。$\{\lambda_{ab\cdots m}\}$ 参数的数量等于从 I 项中有放回地选取 T 项的可能次数，即为 $\binom{I+T-1}{T}$。因此，模型的残差自由度为 $\mathrm{df} = I^T - \binom{I+T-1}{T}$（Haberman，1978：p. 518）。

在一个 I^T 表格中，如果存在

$$\log \mu_{\mathbf{i}} = \lambda_{1i_1} + \lambda_{2i_2} + \cdots + \lambda_{Ti_T} + \lambda_{ab\cdots m}, \tag{10.33}$$

其中 $\lambda_{ab\cdots m}$ 与在完全同质性模型中的定义相同，那么我们说该表格满足*准对称性*（*quasi-symmetry*）。它包括对称性关联以及高阶交互项，但同时允许每个单变量的边际分布具有单独的参数。模型的可识别性条件可以是对每个 t 都设定 $\lambda_{tI} = 0$。在准对称性模型中，其中的一组主效应项是冗余的（习题 10.31）。这个模型比完全对称性模型多 $(I-1)(T-1)$ 个参数，可以通过迭代法来拟合该模型。

当结果变量为定序变量时，对主效应使用排序的赋值 $\{u_a\}$，可以得到更简单的模型。*定序准对称性模型*（*ordinal quasi-symmetry model*）可表示为

$$\log \mu_{\mathbf{i}} = \beta_1 u_{i_1} + \beta_2 u_{i_2} + \cdots + \beta_T u_{i_T} + \lambda_{ab\cdots m},$$

其中，可以设定 $\beta_T = 0$。完全对称性模型是上述模型中 $\beta_1 = \cdots = \beta_T$ 时的一个特例。

当式 10.33 准对称性模型或定序准对称性模型成立时，边际同质性等价于完全对称性。如果准对称性模型（QS）成立但完全对称性模型（S）不成立，那么，这表明存在边际异质性。统计量

$$G^2(\mathrm{S}\mid\mathrm{QS}) = G^2(\mathrm{S}) - G^2(\mathrm{QS})$$

可以用来检验边际同质性。在完全对称性成立时，该统计量渐近服从自由度为 $\mathrm{df} = (I-1)(T-1)$ 的卡方分布。利用定序准对称性模型进行相应检验的自由度为 $\mathrm{df} = (T-1)$。

10.7.2 例子：对合法流产的态度

参见表 10.13。调查对象对是否在以下三种情况下支持合法流产表达了自己的态度：(1)家庭收入太低，无法抚养更多的孩子；(2)怀孕妇女还未婚并且不打算嫁给胎儿的父亲；(3)不论出于何种原因，只要该妇女不想要孩子。表 10.13 按照调查对象的性别将上述结果进行了划分，形成了一个 2^4 表格。

令 μ_{ghij} 表示性别为 g（1 = 女性；0 = 男性）的被访者对三个问题的回答依次为 (h,i,j) 时的期望频数。考虑模型

$$\log \mu_{ghij} = \beta g + \lambda_{abc},$$

表 10.13 按照性别划分的对三种不同情况下进行合法流产的态度

性别	对三个问题的问答结果[a]							
	(1,1,1)	(1,1,2)	(2,1,1)	(2,1,2)	(1,2,1)	(1,2,2)	(2,2,1)	(2,2,2)
男性	342	26	6	21	11	32	19	356
女性	440	25	14	18	14	47	22	457

a 三个问题分别是(1)家庭收入太低而无法抚养更多的孩子;(2)怀孕妇女还未婚并且不打算嫁给胎儿的父亲;(3)不论出于何种原因,只要该妇女不想要孩子。1,支持;2,不支持。

来源:数据取自 1994 年综合社会调查(General Social Survey),美国民意研究中心(National Opinion Research Center)。

其中,当 $(h,i,j) = (1,1,1)$ 时,交互项为 λ_{111};当 $(h,i,j) = (1,1,2)$、$(1,2,1)$ 或 $(2,1,1)$ 时,交互项为 λ_{112};当 $(h,i,j) = (1,2,2)$、$(2,1,2)$ 或 $(2,2,1)$ 时,交互项为 λ_{122};当 $(h,i,j) = (2,2,2)$ 时,交互项为 λ_{222}。这个模型意味着,对每个性别而言,存在完全对称的概率模式。模型的拟合结果为 $G^2 = 39.2$,自由度 df = 11。

加入关于这三个问题的主效应项意味着,对每个性别来说,具有同样的准对称性模式。相应模型对数据的拟合状况明显改进,$G^2 = 10.2$,自由度 df = 9。因此,关于存在对称性的关联结构的假设看上去是合理的。事实上,在只包括二维关联项的对数线性模型中,所拟合的关于问题 1 和问题 2 的对数发生比之比为 3.2,问题 1 和问题 3 的对数发生比之比为 2.6,问题 2 和问题 3 的对数发生比之比为 3.3。

这里,可以通过似然比统计量 $39.2 - 10.2 = 29.0$ 来检验给定性别后的边际同质性,其自由度为 df = 2。对准对称性模型中主效应的分析表明,与其他两种情况相比,在家庭收入很低而无法抚养更多孩子的情况下,对合法流产的支持率更高。

10.7.3 边际对称性的类型

关于 I^T 表格,边际同质性和完全对称性是一般化对称性的特例。在 I^T 表格中,$P(Y_{t_1} = j_1, \cdots, Y_{t_h} = j_h)$ 是一个 h 维的边际概率,其中 h 的取值在 1 到 T 之间;当 $h = 1$ 时,它对应于单变量的边际概率。如果对所有的 h 元组(tuples) $\mathbf{j} = (j_1, \cdots, j_h)$,该边际概率对每个 $\mathbf{j}$ 的排列以及 T 个变量中的所有 h 维组合 $\mathbf{t} = (t_1, \cdots, t_h)$ 都相等,那么我们称该表格满足 h 阶边际对称性(*hth-order marginal symmetry*)。

当 $h = 1$ 时,第一阶边际对称性等价于边际同质性。如果对所有的 t 和 u,$P(Y_t = a, Y_u = b)$ 都相等并且对所有结果配对 (a,b) 都成立,那么我们就称这个表格满足第二阶边际对称性。换句话说,该表所对应的二维边际表格满足对称性,并且都相等。I^T 表格的第 T 阶边际对称性等价于完全对称性。

当第 h 阶对称性成立时,对任意 $i < h$,第 i 阶边际对称性也成立。例如,完全对称性隐含着存在第二阶边际对称性,而第二阶边际对称性又隐含着边际同质性。尽管这种层级关系在数学形式上很有吸引力,但是在现实应用中,高阶对称性模型由于限定过多而很少能给出较好的拟合结果。

10.7.4 边际模型:多维表格

在现实应用中,联合分布的形式通常并不是研究关注的重点。相反,许多研究问题是与边际分布有关的。第 10.3 节的关于配对数据的边际模型可以扩展到匹配集数据的情况。例如,当变量使用定序尺度时,相应的累积 Logit 模型为

$$\text{Logit}[P(Y_t \leqslant j)] = \alpha_i + \beta_t, \quad j = 1, \cdots, I-1, \quad t = 1, \cdots T。 \qquad (10.34)$$

在下一章中,我们将介绍更一般意义上的边际模型,将本章的分析扩展到包括匹配集以及解释变量的情况。

注 解

第 10.1 节:相依比例的比较

10.1 Miettinen(1969)将 McNemar 检验扩展到每个个案包括多个控制案例的个案—控制研究的情况。这时,表 10.2 的表述形式很有意义。n 个匹配集形成了一个 $2 \times 2 \times n$ 表格的层,其中每层的第一列(个案)只有一个观测值,而第二列(控制)包括多个观测值。

Altham(1971)以及 Ghosh 等(2000)介绍了关于二分配对数据的贝叶斯分析。Copas(1973)、Gart(1969)、Kenward 和 Jones(1994)以及 Miettinen(1969)探讨了一般化的配对设计。其中的某些方法(Ghosh et al.,2000;Liang and Zeger,1988;Suissa and Shuster,1991)在推断边际同质性时也利用了主对角线上的观测值。

第 10.4 节:对称性、准对称性以及准独立性

10.2 有关准对称性模型的其他讨论,参见:Darroch(1981)、McCullagh(1982)。术语*准独立性*(*quasi-independence*)是由 Goodman(1968)提出的。它的更一般化的定义为,对于某些固定数量的单元格,存在 $\pi_{ab} = \alpha_a\beta_b$,参见:Caussinus(1966),Fienberg(1970b,1972),Goodman(1968)。Caussinus 利用这一概念分析了删除表中的某组单元格的情况,而 Goodman 在早期关于社会流动的分析中使用了这一模型。Altham(1975)利用它分析了三角形表格,即观测值只落在主对角线以上或以下的单元格的表格。Stigler(1999,chap. 19)综述了相应的早期应用,包括 1913 年 Karl Pearson 对三角形数组的处理。Booth 和 Butler(1999)以及 Smith 等(1996)介绍了有关方形表格模型的精确检验。

10.3 定序准对称性模型的效应 β 与对象别相邻类别 Logit 模型中的场合效应存在联系(Agresti,1993)。条件对称性模型是*对角参数对称性*(*diagonals-parameter symmetry*)模型的一个特例,

$$\log(\pi_{ab}/\pi_{ba}) = \tau_{b-a}, \quad a < b。$$

参见:Goodman(1979b,1985),Hout 等(1987)。

10.4 在某些应用中,表格具有先验(*priori*)的对称性或独立性,但是由于只能观察到配对 (i,j) 而无法观察它们的次序,因而形成一个上三角形表格。有关例子以及对满足层内对称性(symmetric within layers)的三维表格的最大似然拟合,参见:Khamis(1983)。

第 10.5 节:不同评定者之间评定结果的一致性

10.5 卡帕和加权卡帕与组内相关系数有关,后者是在定距尺度下测量评定者之间可靠性的一种指标(Fleiss,1981;Fleiss and Cohen,1973;Kraemer,1979)。Banerjee 等(1999)以及 Fleiss(1981,第 13 章)对卡帕及其发展进行了回顾。有关利用对数线性模型分析一致性的例子,参见:Becker and Agresti(1992)、Goodman(1979b)、Tanner and Young(1985),以及习题 10.41。Darroch 和 McCloud(1986)介绍了准对称性模型在一致性分析中的重要地位。

第 10.6 节:关于成对选择的 Bradley-Terry 模型

10.6 Zermelo(1929)提出了一个与 Bradley-Terry 模型等价的模型。Luce(1959)给出了

有关的理论基础。Mosteller(1951)以及Thurstone(1927)提出了一个使用probit连结的相似模型。在M. Hollander对Ralph Bradley进行的一次有趣采访中(*Stat. Sci.* 16: 75-100,2001),他们讨论了促使该模型发展的关于食物品评的应用。有关更详细的讨论,参见:Bradley(1976)。Fienberg和Larntz(1976)以及Imrey等(1976)探讨了它与准独立性模型的联系。Dittrich等(1998)介绍了在模型中允许协变量的情况。Matthews和Morris(1995)给出了一个涉及析因设计(factorial design)、平分情况(ties)并允许评定间存在相依性的应用。Böckenholt和Dillon(1997)对定序偏好之间的相依性进行了模型分析。David(1988)以及Imrey(1998)综述了有关成对选择的方法。

习题

应用部分

10.1 表10.14显示了调查对象在被问及以下两个问题时的回答结果:"如果一个人得了不治之症,你认为他有权利结束自己的生命吗?"以及"当一个人得了不治之症时,如果病人及其家属申请无痛死亡,你认为应当允许医生这样做吗?"。表中将这两个变量分别称为"自杀"和"安乐死"。

表10.14 习题10.1的数据

自杀	安乐死	
	是	否
是	1 097	90
否	203	435

来源:1994年综合社会调查(General Social Survey),美国民意研究中心(National Opinion Research Center)。

a. 利用置信区间比较它们的边际比例。

b. 对数据进行McNemar检验,并解释结果。

c. 求式10.8模型中关于β的条件最大似然估计。解释结果。

10.2 参考表8.16和习题8.1。将数据视为根据性别进行分层后相应观点的配对数据。对由第1行的(6,160)和第2行的(11,181)所组成的2×2表格进行独立性检验,实际上等于检验每个性别对应的式10.8Logit模型中的β相等。说明其中的原因。

10.3 某项研究对100个研究对象进行交叉实验,以比较两种药品治疗偏头痛的效果。实验结果分别为成功(1)和失败(0)。随机选取的一半研究对象在第一次头痛时使用药品A,然后在下一次头痛时使用药品B。其中,对(A,B)来说,有6个人的结果为(1,1),25人为(1,0),10人为(0,1),并有9人为(0,0)。按照相反次序服用药物的其他50个研究对象,(A,B)的结果如下:10人为(1,1),20人为(1,0),12人为(0,1),以及8人为(0,0)。

a. 忽略服药的次序,比较两种药品的成功概率。解释结果。

b. McNemar检验只使用两个观测值结果不同的配对数据。对于这项研究,表10.15给出了分不同治疗次序的上述配对数据。关于该表的独立性检验是对两种治疗方式的成功率是否相等的检验(Gart,1969)。说明其中的原因。分析这些数据,并解释结果。

表 10.15　习题 10.3 的数据

治疗次序	效果较好的治疗方式	
	第一种	第二种
先 A 后 B	25	10
先 B 后 A	12	20

10.4　一项个案—控制研究包括 8 对研究对象，其中个案患有结肠癌，控制案例则是按照个案的性别和年龄进行匹配的对象。一个潜在的解释变量是研究对象饮食中红肉所占的比重，其测量结果分为“1 = 高”或“0 = 低”。关于（个案，控制）的观测结果分别为，有 3 对取值为（1,1），1 对为（0,0），3 对为（1,0），以及 1 对为（0,1）。

a. 将 8 对个案和控制案例的数据按照饮食（1 或 0）进行交叉划分。我们把这个表称为表 A。给出由 8 个有关饮食情况（1 或 0）与结果变量（个案或控制）之间的分表组成的 2 × 2 × 8 表格，称之为表 B。

b. 计算表 A 的 McNemar z^2 以及表 B 的 CMH 统计量。加以比较。

c. 证明关于表 B 存在共同发生比之比的 Mantel-Haenszel 估计等于表 A 中的 n_{12}/n_{21}。

d. 将表 B 中个案与控制案例的饮食情况取值相同的配对删除，证明这并不影响有关的 CMH 统计量以及对发生比之比的 Mantel-Haenszel 估计。

e. 对大样本检验来说，这项研究的样本规模太小。利用二项分布，以饮食中红肉比重较大的结肠癌的发病率较高为备择假设，求表 A 中的边际同质性检验的精确 P 值。

10.5　《综艺（*Variety*）》杂志每周都会发布几个城市的影评人对新电影的评论情况。每个评论基于总体评价为正面、负面或正负都有而划分为赞同、反对或混合三类。表 10.16 列出了在 1995 年 4 月至 1996 年 9 月期间，芝加哥影评人 Gene Siskel 和 Roger Ebert 的评论情况。

表 10.16　习题 10.5 的数据

Siskel	Ebert		
	反对	混合	赞同
反对	24	8	13
混合	8	13	11
赞同	10	9	64

来源：A. Agresti and L. Winner, *CHANCE* 10:10-14(1997)，经美国统计学会（the American Statistical Association）授权重印，版权 1997。

a. 分别拟合对称性模型、准独立性模型以及准对称性模型。解释结果。

b. 利用模型进行边际同质性检验，并加以解释。

c. 利用一致性模型和/或有关的一致性测量指标分析这些数据。

10.6　参见表 10.5。利用赋值 $u_1 = 1$，$u_4 = 4$，并适当选择不等距的 u_2 和 u_3 的值，拟合定序准对称性模型。将结果与第 10.3.2 节和第 10.4.7 节的结果加以比较，并进行解释。

10.7　参见表 8.19 的所有四个项目。

a. 拟合完全对称性模型和准对称性模型。检验边际同质性，并解释结果。

b. 拟合定序准对称性模型。检验边际同质性,并对模型中的效应加以解释。

10.8 表 10.17 显示了调查对象在两个时点购买不含咖啡因的速溶咖啡的选择。

a. 拟合对称性模型,并利用残差分析两次购买行为的变化。

b. 检验边际同质性。证明 P 值较小反映了选择 High Point 的比例下降了而选择 Sanka 的比例上升了,对其他咖啡的选择没有明显变化。

c. 表明准独立性模型具有 $G^2 = 13.8$ (df = 11)。解释结果,并指出还有哪些分析可能有价值。

表 10.17 习题 10.8 的数据

第一次购买	第二次购买				
	High Point	Taster's Choice	Sanka	Nescafe	Brim
High Point	93	17	44	7	10
Taster's Choice	9	46	11	0	9
Sanka	17	11	155	9	12
Nescafe	6	4	9	15	2
Brim	10	4	12	2	27

来源:数据取自:R. Grover and V. Srinivasan, *J. Market. Res.* 24:139-153(1987)。经 the American Marketing Association 授权重印。

10.9 表 10.18 所示为英国某一样本中父亲职业地位与儿子职业地位之间的关系。利用以下模型分析这一数据:(a)对称性,(b)准对称性,(c)定序准对称性,(d)条件对称性,(e)边际同质性,(f)准独立性,(g)准一致性关联。根据相应的拟合情况解释结果。

表 10.18 习题 10.9 的数据

父亲的地位	儿子的地位					小计
	1	2	3	4	5	
1	50	45	8	18	8	129
2	28	174	84	154	55	495
3	11	78	110	223	96	518
4	14	150	185	714	447	1 510
5	3	42	72	320	411	848
小计	106	489	459	1 429	1 017	3 500

来源:授权重印自:D. V. Glass(ed), *Social Mobility in Britain*, Glencoe, IL: Free Press(1954)。

10.10 对于表 10.18,利用卡帕指标来描述一致性。解释结果。

10.11 表 10.19 显示了两位神经学家对来自两个城市——温尼伯(Winnipeg)和新奥尔良(New Orleans)——的患者进行硬化症诊断的结果。诊断结果的分类为:(1)确定;(2)很可能;(3)有可能;(4)不太可能或绝对不是。对于新奥尔良的患者,利用下列方法分析诊断结果的一致性:(a)独立性模型及其残差,(b)更复杂的模型,(c)卡帕。分别加以解释。

表 10.19 习题 10.11 的数据

新奥尔良神经学家	温尼伯神经学家							
	温尼伯患者				新奥尔良患者			
	1	2	3	4	1	2	3	4
1	38	5	0	1	5	3	0	0
2	33	11	3	0	3	11	4	0
3	10	14	5	6	2	13	3	4
4	3	7	3	10	1	2	4	14

来源:J. R. Landis and G. G. Koch, Biometrics 33:159-174(1977)。经 the Biometric Society 授权重印。

10.12 对于习题 10.11,构建一个同时反映神经学家对两地患者诊断结果的一致性的模型。

10.13 在一个 4 × 4 表格中,对所有 i, 具有 $n_{ii} = 5$; 对 $i = 1,2,3$, $n_{i,i+1} = 15$; $n_{41} = 15$; 其他 $n_{ij} = 0$。计算这个表格的卡帕。说明为什么强关联并不一定就意味着强一致性。

10.14 参见表 10.8。根据模型所给出的标准化残差,说明为什么式 9.6 双线性关联模型可能会对数据拟合得很好。拟合该模型,并描述其中的关联。

10.15 1990 年,佛罗里达大学心理系的研究生参与了一项调查。一个被选中的样本被要求品尝三种没有标识的可乐饮料,并对其进行两两排序。在 49 次对可口可乐与百事可乐的比较中,29 人偏好可口可乐。在 47 次原味可乐与百事可乐的比较中,19 人偏好原味可乐。在 50 次可口可乐与原味可乐的比较中,31 人选择可口可乐。在此,不包括对两种饮料打分相同的情况。

a. 拟合 Bradley-Terry 模型,分析拟合结果的好坏,并对三种饮料进行排序。这里,是否存在充分的证据表明某种饮料最受偏爱?

b. 利用该模型估计对可口可乐的偏好超过百事可乐的概率,并与相应的样本比例进行比较。

10.16 表 10.20 给出了四种统计学期刊在 1987—1989 年期间被引用的情况。期刊上的文章被引用得越多,就表明这个期刊越具权威性。对于涉及两种期刊 A 和 B 的索引,如果是 B 索引了 A, 则视为 A 的一次成功;如果相反,则视为 A 的一次失败。对该数据拟合 Bradley-Terry 模型。解释模型的拟合结果,并给出关于期刊权威性的排序。针对在 *Commun. Stat.* 和 *JRSS-B* 之间发生的索引,估计由 *Commun. Stat.* 引用 *JRSS-B* 的文章的概率。

10.17 表 10.21 显示了几位网球女运动员在 1989—1990 年间的交锋记录。

a. 拟合 Bradley-Terry 模型。解释结果,并对她们进行排序。

b. 估计塞莱斯击败格拉芙的概率。将模型估计的结果与样本比例进行比较。构建关于此概率的 95% 的置信区间。

c. 按照 80% 的 Bonferroni 同步对比(simultaneous Bonferroni comparison),哪一对选手之间存在显著性差异?

10.18 参见关于篮球罚球命中率的习题 3.3。分析这些数据。

10.19 参见表 2.12 以及习题 2.19。通过模型来分析丈夫性享受与妻子性享受之间的关系。

10.20 参见表 8.19。有关对环境(行变量)和城市(列变量)的回答所构成的二维表具有以下分行的单元格计数:(108,179,157/21,55,52/5,6,24)。分析这些数据。

理论与方法

10.21 对以下类比加以说明:二分数据的 McNemar 检验就好比正态分布数据的成对差异性 t 检验。

表 10.20 习题 10.16 的数据

引用期刊	被引用期刊			
	Biometrika	*Commun. Stat.*	*JASA*	*JRSS-B*
Biometrika	714	33	320	284
Commun. Stat.	730	425	813	276
JASA	498	68	1 072	325
JRSS-B	221	17	142	188

来源:Stigler(1994)。经 the Institute of Mathematical Statistics 授权重印。

表 10.21 习题 10.17 的数据

胜方	败方				
	塞莱斯	格拉芙	萨巴蒂尼	纳芙拉蒂诺娃	桑切斯
塞莱斯	—	2	1	3	2
格拉芙	3	—	6	3	7
萨巴蒂尼	0	3	—	1	3
纳芙拉蒂诺娃	3	0	2	—	3
桑切斯	0	1	2	1	—

10.22 对于 2 × 2 表格,推导 $\text{cov}(p_{+1}, p_{1+})$,并证明 $\text{var}[\sqrt{n}(p_{+1} - p_{1+})]$ 等于式 10.1。

10.23 参考关于二分配对数据的对象别式 10.8 模型。

a. 证明 $\exp(\beta)$ 是观测场合与结果之间的条件发生比之比。说说它与式 10.6 模型中的发生比之比 $\exp(\beta)$ 的区别。

b. 利用式 10.9 条件分布,证明 $\hat{\beta} = \log(n_{21}/n_{12})$。

c. 对于由 n 个配对形成的一个随机样本,说明为什么

$$E(n_{21}/n) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\exp(\alpha_i)}\frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}。$$

类似地,给出 $E(n_{12}/n)$ 的表达式。在给定 n 的情况下,利用当 $n \to \infty$ 时 $E(n_{21}/n)$ 和 $E(n_{12}/n)$ 之间的比率,说明为什么 $n_{21}/n_{12} \to \exp(\beta)$ (提示:应用由 A. A. Markov 提出的关于独立、非同分布的随机变量的大数定律,或者应用切比雪夫不等式(Chebyshev's inequality))。

d. 证明在 $2 \times 2 \times n$ 形式的数据中,关于共同发生比之比的 Mantel-Haenszel 估计值式 6.7 简化为 $\exp(\hat{\beta}) = n_{21}/n_{12}$。

e. 利用 δ 方法,证明 $\hat{\beta}$ 的标准误等于式 10.10。

f. 对于形如表 10.2 的表格,证明它的式 6.6CMH 统计量在代数上等于形如表 10.1 的相应表格的 McNemar 统计量 $(n_{21} - n_{12})^2/(n_{21} + n_{12})$。

10.24 参考习题 10.23。与 β 的条件最大似然估计不同,关于它的无条件最大似然估计不具有一致性(Anderson,1980:244-245;他最早在 1973 年提出)。按照下列要求对此进行证明:

a. 假定不同对象之间以及同一对象在不同观测场合的取值是独立的,求对数似

然函数。证明似然方程为 $y_{+t} = \sum_i P(Y_{it} = 1)$ 和 $y_{i+} = \sum_t P(Y_{it} = 1)$。

b. 将 $\exp(\alpha_i)/[1+\exp(\alpha_i)] + \exp(\alpha_i+\beta)/[1+\exp(\alpha_i+\beta)]$ 代入第二个似然方程，证明对于 $y_{i+} = 0$ 的 n_{22} 个对象有 $\hat{\alpha}_i = -\infty$，对于 $y_{i+} = 2$ 的 n_{11} 个对象有 $\hat{\alpha}_i = \infty$，对于 $y_{i+} = 1$ 的 $n_{21} + n_{12}$ 个对象有 $\hat{\alpha}_i = -\hat{\beta}/2$。

c. 将 $\sum_i P(Y_{it} = 1)$ 按照对象所对应的 $y_{i+} = 0$，$y_{i+} = 2$ 以及 $y_{i+} = 1$ 分割为几项，证明在 $t = 1$ 时，第一个似然方程为 $y_{+1} = n_{22}(0) + n_{11}(1) + (n_{21} + n_{12})\exp(-\hat{\beta}/2)/[1+\exp(-\hat{\beta}/2)]$。说明为什么 $y_{+1} = n_{11} + n_{12}$，求解第一个似然方程并证明 $\hat{\beta} = 2\log(n_{21}/n_{12})$。因而，根据习题 10.23 的一个结论，$\hat{\beta} \xrightarrow{P} 2\beta$。

10.25 考虑当 Y_1 和 Y_2 相互独立时的式 10.6 边际模型与当 $\{\alpha_i\}$ 都相等时的式 10.8 条件模型。说明为什么它们是等价的。

10.26 令 $\hat{\beta}_M = \log(p_{+1}p_{2+}/p_{+2}p_{1+})$ 表示式 10.6 边际模型的参数估计，$\hat{\beta}_C = \log(n_{21}/n_{12})$ 表示式 10.8 条件模型的参数估计。利用 δ 方法，证明 $\sqrt{n}(\hat{\beta}_M - \beta_M)$ 的渐近方差为

$$(\pi_{1+}\pi_{2+})^{-1} + (\pi_{+1}\pi_{+2})^{-1} - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})。$$

在前面习题中的独立性条件下，$\beta_M = \beta_C$，证明在这种情况下，渐近方差满足

$$\begin{aligned}\mathrm{var}[\sqrt{n}(\hat{\beta}_M)] &= (\pi_{1+}\pi_{2+})^{-1} + (\pi_{+1}\pi_{+2})^{-1} \\ &\leqslant (\pi_{1+}\pi_{+2})^{-1} + (\pi_{+1}\pi_{2+})^{-1} \\ &= \pi_{12}^{-1} + \pi_{21}^{-1} = \mathrm{var}[\sqrt{n}(\hat{\beta}_C)]。\end{aligned}$$

10.27 参考关于配对研究的式 10.12 模型，证明在条件最大似然法中，条件分布满足式 10.13，并且当 $S_i = 0$ 或 2 时它不取决于 $\boldsymbol{\beta}$。指出当一个预测变量对所有 i 都存在 $x_{ji1} = x_{ji2}$ 时，条件分布中的 β_j 会出现什么情况。

10.28 考虑包括 T 个观测值的匹配集数据而不是配对数据的式 10.12 模型，说明如何对式 10.13 进行扩展，并构建它的条件似然函数。

10.29 举例说明，当 $I > 2$ 时，边际同质性并不意味着对称性。

10.30 推导下列模型的似然方程和残差自由度：(a) 对称性，(b) 准对称性，(c) 准独立性，(d) 定序准对称性。

10.31 在式 10.19 准对称性模型中，令 $\lambda_a = \lambda_a^X - \lambda_a^Y$，证明可以将其等价地表示为 $\log\mu_{ab} = \lambda + \lambda_a + \lambda_{ab}^*$，其中 $\lambda_{ab}^* = \lambda_{ba}^*$，也即，这里只需要一组主效应参数。

10.32 证明准对称性等价于下式（Caussinus, 1966），即对所有 a, b 和 c，存在

$$(\pi_{ab}\pi_{bc}\pi_{ca})/(\pi_{ba}\pi_{cb}\pi_{ac}) = 1。$$

10.33 推导关于差异向量（difference vector）$\mathbf{d}$ 的式 10.16 协方差矩阵。

10.34 构建一个既满足边际同质性，又满足统计独立性的对数线性模型。证明 $\hat{\pi}_{ab} = (p_{+a} + p_{a+})(p_{+b} + p_{b+})/4$，并且残差自由度为 $\mathrm{df} = I(I-1)$。

10.35 考虑式 10.28 条件对称性（CS）模型。

a. 证明它的对数线性表达形式为

$$\log\mu_{ab} = \lambda_{\min(a,b),\max(a,b)} + \tau I(a < b),$$

其中 $I(\cdot)$ 是一个指示函数（另见：Bishop et al., 1975: 285-286）。

b. 证明似然方程为：

$$对所有\ a \leqslant b,\quad \hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba},\quad \sum_{a<b}\sum \hat{\mu}_{ab} = \sum_{a<b}\sum n_{ab}。$$

c. 证明 $\hat{\tau} = \log\left[\left(\sum\sum_{a<b} n_{ab}\right)/\left(\sum\sum_{a>b} n_{ab}\right)\right]$；$\hat{\mu}_{aa} = n_{aa}$，$a = 1,\cdots,I$；对于 $a \neq b$，有 $\hat{\mu}_{ab} = \exp[\hat{\tau} I(a < b)](n_{ab} + n_{ba})/[\exp(\hat{\tau}) + 1]$。

d. 证明关于 $\hat{\tau}$ 的渐近方差的估计值等于

$$\left(\sum_{a<b}\sum n_{ab}\right)^{-1} + \left(\sum_{a>b}\sum n_{ab}\right)^{-1}。$$

e. 证明残差自由度为 $df = (I+1)(I-2)/2$。

f. 证明条件对称性 + 边际同质性 = 对称性。说明为什么 $G^2(S|CS)$ 是关于边际同质性的检验(df=1)。当模型成立时，$G^2(S|CS)$ 在渐近意义上比 $G^2(S|QS)$ 具有更强的统计效能。为什么？

10.36 指出下列 Logit 模型所对应的对数线性模型。对于 $a < b$，$\log(\pi_{ab}/\pi_{ba})$ 等于：(a)0，(b) τ，(c) $\alpha_a - \alpha_b$，(d) $\beta(b-a)$。

10.37 一个不基于模型的边际异质性的定序测量指标为

$$\hat{\Delta} = \sum_{a<b}\sum p_{a+}p_{+b} - \sum_{a>b}\sum p_{a+}p_{+b}。$$

证明 $\hat{\Delta}$ 是对 $\Delta = P(Y_1 > Y_2) - P(Y_2 > Y_1)$ 的估计，其中 Y_1 具有分布 $\{\pi_{a+}\}$ 并且 Y_2 独立于 $\{\pi_{+b}\}$。证明边际同质性意味着 $\Delta = 0$。证明关于 $\hat{\Delta}$ 的渐近方差的估计为

$$\left[\sum_a\sum_b \hat{\phi}_{ab}^2 p_{ab} - \left(\sum_a\sum_b \hat{\phi}_{ab} p_{ab}\right)^2\right]/n,$$

其中，$\hat{\phi}_{ab} = \hat{F}_{b1} + \hat{F}_{b-1,1} - \hat{F}_{a2} - \hat{F}_{a-1,2}$，$\hat{F}_{a1} = (p_{1+} + \cdots + p_{a+})$，$\hat{F}_{a2} = (p_{+1} + \cdots + p_{+a})$ (Agresti，1984：208-209)。

10.38 对于排序的赋值 $\{u_a\}$，令 $\bar{y}_1 = \sum_a u_a p_{a+}$，$\bar{y}_2 = \sum_a u_a p_{+a}$，证明边际同质性意味着 $E(\bar{Y}_1) = E(\bar{Y}_2)$，并且

$$\left[\sum_a\sum_b (u_a - u_b)^2 p_{ab} - (\bar{y}_1 - \bar{y}_2)^2\right]/n$$

是对 $\mathrm{var}(\bar{Y}_1 - \bar{Y}_2)$ 的估计。构建一个关于边际同质性的检验(Bhapkar，1968)。

10.39 考虑关于方形表格的可积模型，

$$\pi_{ab} = \begin{cases} \alpha_a\alpha_b(1-\beta), & a \neq b \\ \alpha_a^2 + \beta\alpha_a(1-\alpha_a), & a = b \end{cases}。$$

a. 证明该模型满足：(i)对称性，(ii)边际同质性，(iii)准对称性，(iv)准独立性。

b. 证明 $\alpha_a = \pi_{a+} = \pi_{+a}$，$a = 1,\cdots,I$。

c. 证明 β = Cohen 的卡帕，并对该模型在 $\kappa = 0$ 和 $\kappa = 1$ 时进行解释。

10.40 一个 2×2 表格的真实发生比之比等于 10。求满足以下条件的单元格概率：(a) $\pi_{1+} = \pi_{+1} = 0.5$，(b) $\pi_{1+} = \pi_{+1} = 0.3$，(c) $\pi_{1+} = \pi_{+1} = 0.1$。求每种情况下的卡帕值(这一结果表明，对于某个给定的关联，卡帕严重依赖于边际概率的大小；另见：Sprott(2000：59))。

10.41 结果变量为定序变量的一致性模型将随机因素无法解释的一致性分割为两个部分：基线的关联和主对角线增量(A. Agresti，*Biometrics* 44：539-548，1988)。对于排序的赋值 $\{u_a\}$，该模型为

$$\log\mu_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \beta u_a u_b + \delta I(a = b)。\quad (10.35)$$

a. 证明这是准对称性模型和准关联模型式 10.29 的一个特例。

b. 对于一致性发生比式 10.30,证明 $\tau_{ab} = (u_b - u_a)^2\beta + 2\delta$。利用具有单位距离的赋值,证明由四个主对角线以外的单元格形成的局部发生比之比满足 $\log \theta_{ab} = \beta$。

c. 求它的似然方程,并证明 $\{\hat{\mu}_{ab}\}$ 和 $\{n_{ab}\}$ 具有相同的边际分布、相关系数,以及完全一致情况的发生比例。

d. 在表 10.8 中,利用 $\{u_a = a\}$ 指出,式 10.35 模型的 $G^2 = 4.8$ (df = 7), $\hat{\delta} = 0.842(SE = 0.427)$, $\hat{\beta} = 1.316(SE = 0.420)$。利用 $|a - b| > 1$ 时的 $\hat{\tau}_{a,a+1}$ 和 $\hat{\theta}_{ab}$ 对结果加以解释。

10.42 参考 Bradley-Terry 模型。

a. 证明 $\log(\Pi_{ac}/\Pi_{ca}) = \log(\Pi_{ab}/\Pi_{ba}) + \log(\Pi_{bc}/\Pi_{cb})$。

b. 在这个模型中,有没有可能对 a 的偏好大于对 b 的偏好(即 $\Pi_{ab} > \Pi_{ba}$),对 b 的偏好大于对 c 的偏好,而对 c 的偏好大于对 a 的偏好? 加以说明。

c. 说明为什么如果不设定诸如 $\beta_I = 0$ 的约束条件,那么模型中的 $\{\beta_a\}$ 就无法唯一确定(提示:证明对于任意 c, 当 $\{\beta_a^* = \beta_a - c\}$ 时,模型都成立)。

10.43 参考式 10.32 模型。

a. 构建一个包括主队参数 $\{\beta_{Hi}\}$ 和客队参数 $\{\beta_{Ai}\}$ 的更一般的模型,使得当 i 为主队时,球队 i 击败球队 j 的概率为 $\exp(\beta_{Hi})/[\exp(\beta_{Hi}) + \exp(\beta_{Ai})]$, 其中 $\beta_{AI} = 0$ 但对 β_{Hi} 的取值不加限定。

b. 在下列情况下对 $\{\beta_{Hi} = \beta_{Ai} + c\}$ 进行解释:(i) $c = 0$, (ii) $c > 0$。

c. 对表 10.12 的数据拟合该模型。将其结果与式 10.32 模型的拟合情况进行比较。通过对比 $\{\hat{\beta}_{Hi}\}$ 和 $\{\hat{\beta}_{Ai}\}$ 来描述球队在主客场的表现。

10.44 求 Bradley-Terry 模型的对数似然函数。根据其核函数,证明(给定 $\{N_{ab}\}$)最小充分统计量是 $\{n_{a+}\}$, 并解释"总胜场数"如何决定了所估计的排序。

10.45 说明如何对 T 维数据拟合完全对称性模型。

10.46 证明如果第 k 阶边际对称性成立,那么对于任意 $j < k$, 第 j 阶边际对称性都成立。

10.47 假定一个 I^T 表格满足准对称性模型,对该表按照其中的一个变量进行合并后,证明该模型对所得到的 I^{T-1} 表格仍然成立,并且其主效应保持不变。

11 对重复测量的分类结果变量的分析

许多研究都需要在多个时点或在不同情况下对研究对象的结果变量进行重复测量。重复测量的分类结果变量在与健康有关的应用中非常普遍，尤其是在跟踪研究中。例如，医生可能会每隔一周就对病人进行检查，以评估某种新药品的治疗效果。在一些情况下，解释变量的取值可能会随时间而发生变化。当然，重复测量的结果变量并不局限于不同的时点，比如一项牙科研究可能会测量研究对象口腔中的每颗牙齿是否存在蛀牙的问题。

一般情况下，这类结果变量对应的是研究对象的*匹配集*（*matched sets*）或*群组*（*clusters*）。例如，对怀孕母鼠的样本使用不同剂量的毒素，观察一窝子鼠中每个胎儿的存活状况（存活，死亡）。再如，在一项使用多阶段抽样设计分析儿童肥胖症的影响因素的研究中，可以将属于同一家庭的孩子作为一个群组。同一群组内的观测值往往比来自不同群组间的观测值彼此更相似。如果忽略了这种群组现象，用普通方法来分析群组数据会导致严重的问题。

在本章中，我们对第 10 章介绍的关于配对数据的分析方法进行扩展。第 11.1 节介绍关于 T 维表格的边际分布的比较。本章剩下的几节讨论在模型中包括解释变量的情况。例如，许多研究需要比较不同组之间或在不同干预方式下的重复测量结果。在第 11.2 节，我们利用最大似然法来拟合边际模型。第 11.3 节介绍*广义估计方程*（*generalized estimating equations*，*GEE*），该方法是类似然法的一种多元形式，它在运算上比最大似然法更为简便。第 11.4 节讨论有关 GEE 方法的技术细节。在本章的最后一节中，我们介绍一种根据前期结果来对观测值进行分析的*转换*（*transitional*）模型。

11.1 边际分布的比较：多元结果变量的情况

一般来说，对于重复测量的多元结果变量，与其多变量相依性相比，我们更关注这些变量的边际分布。例如，在对某种慢性病（如恐惧症）进行治疗时，研究的首要目的是分析在 T 周的治疗期间成功的概率是否会增加。这 T 个成功概率指的是 T 个一阶边际分布。在第 10.2.1 节和第 10.3 节中，我们利用有关边际分布的模型分析比较了配对数据（$T = 2$）的边际分布。本节将这种方法扩展到 $T > 2$ 的情况。

11.1.1 二分变量的边际模型与边际同质性

将 T 个二分结果变量表示为（$Y_1, Y_2, \cdots, Y_T$）。对配对数据的边际 Logit 模型式 10.6

进行扩展可得：

$$\text{Logit}[P(Y_t = 1)] = \alpha + \beta_t, \quad t = 1, \cdots, T, \tag{11.1}$$

模型具有诸如 $\beta_T = 0$ 或 $\alpha = 0$ 的参数约束条件。对于可能出现的结果序列 $\mathbf{i} = (i_1, i_2, \cdots, i_T)$，其中每个 $i_t = 0$ 或 1，令

$$\pi_{\mathbf{i}} = P(Y_1 = i_1, Y_2 = i_2, \cdots, Y_T = i_T)。$$

令 $\boldsymbol{\pi}$ 表示结果变量的可能取值 $\mathbf{i}$ 对应的概率向量，它所对应的是对 T 个结果变量进行交叉划分的一个 2^T 表格，$\boldsymbol{\pi}$ 描述了 $(Y_1, Y_2, \cdots, Y_T)$ 的联合分布。样本单元格比例是关于 $\boldsymbol{\pi}$ 的最大似然估计，并且有关 $y_t = 1$ 的样本比例是对 $P(Y_t = 1)$ 的最大似然估计。

式 11.1 模型是一个饱和模型，它利用 T 个参数描述了 T 个边际概率。边际同质性，即 $P(Y_1 = 1) = \cdots = P(Y_T = 1)$，是模型中 $\beta_1 = \cdots = \beta_T$ 时的一个特例。尽管边际同质性模型只包括一个参数，但是对它进行最大似然拟合并不容易。多项分布似然函数针对的是 2^T 个联合单元格概率 $\boldsymbol{\pi}$，而不是 T 个边际概率 $\{P(Y_t = 1)\}$。具体的拟合方法，详见在第 11.2.5 节。

令 $n_{\mathbf{i}}$ 表示单元格 $\mathbf{i}$ 的样本计数。对数似然函数 $L(\boldsymbol{\pi})$ 的核函数为 $\sum_{\mathbf{i}} n_{\mathbf{i}} \log \pi_{\mathbf{i}}$。令 $L(\mathbf{p})$ 表示对数似然函数在样本比例 $\{p_{\mathbf{i}} = n_{\mathbf{i}}/n\}$ 处的取值，即关于式 11.1 模型的最大似然拟合值。令 $L(\hat{\boldsymbol{\pi}}^{MH})$ 表示在边际同质性假定下对数似然函数的最大值，那么，关于边际同质性的似然比检验（Lipsitz et al.，1990；Madansky，1963）使用以下统计量：

$$-2[L(\hat{\boldsymbol{\pi}}^{MH}) - L(\mathbf{p})] = 2\sum_{\mathbf{i}} n_{\mathbf{i}} \log(p_{\mathbf{i}}/\hat{\pi}_{\mathbf{i}}^{MH})。 \tag{11.2}$$

由于一般模型式 11.1 比边际同质性模型多 $T-1$ 个参数，该统计量渐近服从自由度为 $df = T - 1$ 的卡方零分布。

11.1.2 例子：药品的交叉比较

表 11.1 取自一项交叉研究，该研究在对一种慢性病进行的三次治疗中患者每次服用不同的药物。结果变量将每次治疗的效果分为良好或者不好。这个 2^3 表格的第一个维度列出了对药品 A 的反应（良好，不好），第二个维度给出了对药品 B 的反应，第三个维度为对药品 C 的反应。在这里，我们假定每种药物都不存在延滞效应（carryover effects），并且在整个研究过程中每个研究对象的疾病严重程度没有发生变化。对于许多慢性病比如偏头痛来说，这些假定是合理的。

表 11.1　交叉研究中三种药品的疗效

	对药品 A 反应良好		对药品 A 反应不好	
	对药品 B 反应良好	对药品 B 反应不好	对药品 B 反应良好	对药品 B 反应不好
对药品 C 反应良好	6	2	2	6
对药品 C 反应不好	16	4	4	6

来源：经 the Biometric Societies 授权重印（Grizzle et al.，1969）。

对于药品（A，B，C）反应良好的样本比例为（0.61，0.61，0.35）。相应边际同质性检验的似然比统计量为 5.95（df = 2），对应的 P 值为 0.05。对于同时比较每对药品并保证总体误差不超过 0.05 的置信区间，Bonferroni 法对每个区间应用的置信系数为 $(1 - 0.05/3) = 0.983\ 3$。例如，根据式 10.1，所估计的药品 A 和药品 C 之差为 $0.261 = 0.609 - 0.348$，标准误为 0.108。也就是说，关于两种药品的真实差异的置信区间等于 $0.261 \pm 2.39(0.108)$，即（0.002，0.520）。比较药品 B 和 C 的置信区间与此相同。在这

项研究中,存在一定的证据表明,对药品 C 出现良好反应的比例相对较低。

然而,考虑到样本规模比较小,我们应当审慎对待这些结果。对于每一对药品,可以通过一个 2×2 表格反映两个结果变量之间的关系。相应的 2×2 表格的二项分布精确检验(第 10.4.1 节)使用的是表中主对角线以外的单元格计数。精确检验的结果为,比较药品 A 和 B 的 P 值等于 1.0,比较药品 A 和 C 以及药品 B 和 C 的 P 值都等于 0.036。

11.1.3 对多类别结果变量边际分布的模型分析

二分的边际模型式 11.1 可以扩展到多项分布结果变量的情况,利用结果变量的 I 个类别的基线类别 Logit,相应的饱和模型为

$$\log[P(Y_t = j)/P(Y_t = I)] = \beta_{tj},\quad t = 1,\cdots,T,\quad j = 1,\cdots,I-1。\qquad (11.3)$$

边际同质性模型,即对于 $j = 1,\cdots,I-1$ 都有 $P(Y_1 = j) = \cdots = P(Y_T = j)$,是饱和模型中

$$\beta_{1j} = \beta_{2j} = \cdots = \beta_{Tj},\quad j = 1,\cdots,I-1$$

时的一个特例。通过比较两个模型,相应的检验边际同质性的似然比统计量具有式 11.2 的形式,其自由度为 $df = (T-1)(I-1)$。

当结果变量为定序变量时,可以通过比边际同质性模型更复杂的非饱和模型分析 T 个边际分布的上下波动,其中一个这样的模型为

$$\text{Logit}[P(Y_t \leqslant j)] = \alpha_j + \beta_t,\quad t = 1,\cdots,T,\quad j = 1,\cdots,I-1,\qquad (11.4)$$

模型以诸如 $\beta_T = 0$ 为约束条件。边际同质性模型是上述模型中 $\beta_1 = \cdots = \beta_T$ 时的一个特例,相应检验的自由度为 $df = T-1$。由于累积概率之间存在排序,这里的 $\{\alpha_j\}$ 满足 $\alpha_1 < \cdots < \alpha_{I-1}$。这些模型可以通过最大似然法来拟合,详见第 11.2.5 节。

11.1.4 边际同质性的沃尔德检验和广义 CMH 计分检验

在本章中,我们关注对边际分布的模型分析,而不仅仅是边际同质性检验。不过,除似然比检验以外,还有其他的检验方式可以用来检验边际同质性,在此我们简要介绍其中的两种。

记 $p_j(t)$ 为结果变量 Y_t 落在类别 j 的样本比例,令

$$\bar{p}_j = \sum_t p_j(t)/T,\qquad d_j(t) = p_j(t) - \bar{p}_j,$$

并令 $\mathbf{d}$ 表示关于 $\{d_j(t), t = 1,\cdots,T-1, j = 1,\cdots,I-1\}$ 的向量。令 $\hat{\mathbf{V}}$ 表示关于 $\sqrt{n}\mathbf{d}$ 的协方差矩阵的估计。Bhapkar(1973)提出了一般情形下的沃尔德统计量

$$W = n\mathbf{d}'\hat{\mathbf{V}}^{-1}\mathbf{d}。\qquad (11.5)$$

该统计量是对式 10.16 的扩展,它服从自由度 $df = (I-1)(T-1)$ 的大样本卡方分布。

另外一个统计量是广义 Cochran-Mantel-Haenszel(CMH)统计量的特例(第 7.5.3 节)。回顾关于二分变量($I = 2$)的配对数据($T = 2$)的情况,CMH 统计量可以应用于每层都显示两个结果变量的三维表格(如表 10.2),其中每个研究对象构成一个层。表 10.2 可以扩展成一个包括 n 层的 $T\times I$ 表格,其中第 k 层给出了关于对象 k 的 T 个结果变量的取值。在每层的第 t 行中,对应观测值 t 所处的那一列的值为 1,所有其他列的值为 0(或者,如果该观测值缺失时,所有列的值都为 0)。这种按照对象进行分层的概率分布与对象别模型存在自然的联系,如式 10.8 Logit 模型。但是,这个三维表格的条件独立性(给定对象)对应着在 I^T 表格中变量之间具有可交换性(exchangeability),进而满足边际同质性。在 $T\times I\times n$ 表中,有关条件独立性的广义 CMH 检验同时也是对在较强的可交

换性条件下生成的样本分布的边际同质性检验(Darroch,1981)。用来检验 T 个均值间的变动性的广义 CMH 统计量,对于结果变量为具有固定赋值的定序变量同样适用。

当 $I=2$ 且 $T=2$ 时,这种 CMH 方法与 McNemar 统计量等价。当 $I=2$ 但 $T>2$ 时,将 T 个结果变量视为定类变量的广义 CMH 统计量与 Cochran(1950)所提出的一个统计量相同。他的统计量被称为 Cochran 的 Q(*Cochran's Q*),其自由度为 $df=T-1$(习题 11.22)。

11.2 边际模型:最大似然法

以上分析都是关于边际分布之间的比较,并没有考虑解释变量。接下来,我们引入预测变量。在本节中,我们首先介绍相应的最大似然法,有关模型拟合的细节将留到本节的最后来讨论。

11.2.1 例子:精神抑郁的跟踪研究

在本章及下一章中,我们将通过表 11.2 的数据来展示各种分析方法。该表取自一项比较一种新药与标准药品在治疗精神抑郁方面效果的跟踪研究(Koch et al.,1977)。按照初诊时抑郁的严重程度(轻度或者重度),将研究对象划分为两组。在每组中,通过随机分配,每个研究对象服用两种药品中的一种。在接受治疗后的第 1 周、第 2 周和第 4 周,分别对每个研究对象都进行了复查,诊断结果划分为正常或者不正常。

表 11.2 按照抑郁症初诊情况和治疗方式对三次复查结果的交叉划分

初诊情况	治疗方式	三次复查的结果[a]							
		NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA
轻度	标准药品	16	13	9	3	14	4	15	6
	新药品	31	0	6	0	22	2	9	0
重度	标准药品	2	2	8	9	9	15	27	28
	新药品	7	2	5	2	31	5	32	6

a N,正常;A,不正常。

来源:经 the Biometric Society 授权重印(Koch et al.,1977)。

表 11.2 按照两个解释变量——治疗方式和初诊情况将研究对象分成了四组。由于该研究对二分结果变量(抑郁状况)在 $T=3$ 次的不同时点分别进行了观测,在表 11.2 中,每组都对应着一个 2^3 表格。这三次对抑郁状况的复诊构成了多元结果变量,它共包括三项,其中 Y_t 在结果为正常时等于 1、不正常时等于 0。对四组对象中每组都进行三次重复测量共形成了 12 个边际分布。

令 s 表示初诊时的严重程度,其中 $s=1$ 为严重,$s=0$ 为轻度;d 表示所使用的药品,$d=1$ 为新药品,$d=0$ 为普通药品;t 表示观测的时点。Koch 等(1977)指出,如果时间维度反映了累积的用药剂量,那么 Logit 通常会与时间的对数呈线性关系。因此,他们利用周数(1,2,4)的以 2 为底数的对数,即(0,1,2),作为对时间的赋值。

表 11.3 给出了 12 个边际分布中复诊结果为正常(即 $y_t=1$)的样本比例。例如,根据表 11.2,对于初诊情况为轻度并服用标准药品的研究对象,在一周后结果为正常的样本比例等于 $(16+13+9+3)/(16+13+9+3+14+4+15+6)=0.51$。由表11.3可见,结果为正常的样本比例:(1)在每组中都随着时间的推移而上升;(2)给定初诊情况,在服用新药的病人中,其样本比例的上升速度快于服用标准药品的上升速度;(3)给定观

测时间和治疗方式，初诊情况为轻度的样本比例高于初诊情况为重度的相应比例。在这类研究中，研制新药的公司会希望分析结果能显示服用新药的病人病情改善速度明显更快。

表 11.3 表 11.2 的抑郁数据中结果为正常的样本边际比例

初诊情况	治疗方式	样本比例		
		第 1 周	第 2 周	第 4 周
轻度	标准药品	0.51	0.59	0.68
	新药品	0.53	0.79	0.97
重度	标准药品	0.21	0.28	0.46
	新药品	0.18	0.50	0.83

边际 Logit 模型

$$\text{Logit}[P(Y_t = 1)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t$$

包括解释变量（初诊情况和服药种类）的主效应，以及代表多元结果变量中不同项的变量（时间）的主效应，其中时间的线性效应 β_3 对于每组来说都相等。

对于每组由三个结果变量形成的 2^3 交叉列联表中的八个单元格，一种自然的假定是它们来自于一个多项分布样本，并且在四组之间相互独立。然而，在多项分布似然函数的乘积中，该模型所针对的是 12 个边际概率（2 种药品 × 2 类初诊情况 × 3 个时点）而不是 $4 \times 2^3 = 32$ 个单元格概率。每组中三个二项分布变量的边际概率并不相互独立。最大似然估计需要通过某种迭代程序来最大化多项分布似然函数的乘积，同时还必须满足模型中有关边际概率的约束条件。相应的算法将在第 11.2.5 节介绍。

通过比较表 11.2 中的 32 个单元格计数以及它们的最大似然拟合值，可以用来检查模型的拟合情况。由于该模型利用四个参数来描述 12 个边际 Logit，它的残差自由度为 df = 8。模型拟合的偏离度 $G^2 = 34.6$，它对数据的拟合很差。究其原因，该模型假定病情改善的速率都等于 β_3，但是样本比例清楚地显示，服用新药品时病情改善得更快。

更为现实的模型是允许每种药品的时间效应存在差异，即

$$\text{Logit}[P(Y_t = 1)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt。$$

这个模型所估计的关于标准药品（$d = 0$）的时间效应为 $\hat{\beta}_3 = 0.48$（SE = 0.12），新药品（$d = 1$）的时间效应为 $\hat{\beta}_3 + \hat{\beta}_4 = 1.49$（SE = 0.14）。因而，新药品的斜率比标准药品高出 $\hat{\beta}_4 = 1.01$（SE = 0.18），这表明新药品对病情改善的速度明显更快。该模型的拟合结果要好得多，它的 $G^2 = 4.2$（df = 7）。与更简单的模型相比，G^2 的变动值 $34.6 - 4.2 = 30.4$ 是关于 $H_0: \beta_4 = 0$（即每种药品的时间效应相等）的似然比检验。

模型所估计的关于初诊情况的效应为 $\hat{\beta}_1 = -1.29$（SE = 0.14）；对于每一个药品-时点的取值组合，当初诊情况为重度时，所估计的复查结果为正常的发生比是初诊情况为轻度的相应发生比的 $\exp(-1.29) = 0.27$ 倍。$\hat{\beta}_2 = -0.06$（SE = 0.22）则表明，在一周后（即 $t = 0$ 时）两种药品之间不存在显著差异。在时点 t，所估计的服用新药品后检查结果为正常的发生比是服用标准药品的相应发生比的 $\exp(-0.06 + 1.01t)$ 倍。总之，初诊情况、治疗方式以及时间对复诊结果为正常的概率都具有显著影响。

11.2.2 重复测量的多项分布结果变量的模型分析

关于重复测量的二分结果变量的边际分布模型可以扩展到结果变量包括多个类别

的情况。在观测点 t，结果变量的边际分布具有 $I-1$ 个 Logit。当结果变量为定类变量时，基线类别 Logit 模型可以描述每个结果类别与基线类别之间的发生比。当结果变量为定序变量时，可以考虑使用累积 Logit 模型。

对于某一特定的边际 Logit，模型具有如下的形式

$$\text{Logit}_j(t) = \alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_t, \quad j = 1,\cdots,I-1, \quad t = 1,\cdots。$$

如果结果变量为定序变量，可能会有 $\text{Logit}_j(t) = \text{Logit}[P(Y_t \leqslant j)]$。这时，$\boldsymbol{\beta}_j$ 可以简化为 $\boldsymbol{\beta}$，模型具有比例发生比的形式，即对每个 Logit 的效应都相同。$\boldsymbol{\beta}$ 中的一些参数可能对应着具有下标 t（即时间）的变量，以表示重复的测量。那么，我们可以比较在 $\mathbf{x}$ 的某个特定取值处的边际分布，或者估计 $\mathbf{x}$ 对结果变量的效应。无论在哪种情况下，检查是否存在交互效应都是很关键的。例如，在每个时点 t 处，$\mathbf{x}$ 的效应是不是都相等？

11.2.3 例子：失眠症

表 11.4 所示为一项随机的双盲临床试验的结果，在该试验中，患有失眠症的病人分别服用一种现行安眠药和安慰剂。结果变量是病人自报的入睡时间（以分钟为单位）。每位病人分别在治疗开始前以及治疗持续二周后做了回答。两种干预方式，即安眠药和安慰剂，构成了一个二分的解释变量。分别接受两种干预方式的病人是相互独立的样本。

表 11.4 按照干预方式和调查时点划分的入睡时间

干预方式	基　期	入睡时间			
		跟踪期			
		<20	20～30	30～60	>60
现行药	<20	7	4	1	0
	20～30	11	5	2	2
	30～60	13	23	3	1
	>60	9	17	13	8
安慰剂	<20	7	4	2	1
	20～30	14	5	1	0
	30～60	6	9	18	2
	>60	4	11	14	22

来源：取自：S. F. Francom, C. Chuang-Stein, and J. R. Landis, Statist. Med. 8:571-582(1989)。经 John Wiley & Sons Ltd. 授权重印。

表 11.5 显示了按照干预方式和调查时点划分的四个样本边际分布。从研究基期到随访时，两组的入睡所需时间看上去都下降了，而使用安眠药的一组似乎下降得更多。这表明，可能存在某种交互效应。这里的结果变量是一个连续变量的离散分组，因而，根据第 7.2.3 节的推导，使用累积连结模型是一种自然的选择。比例发生比模型可表示为

$$\text{Logit}[P(Y_t \leqslant j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 tx \tag{11.6}$$

它允许在 t = 时点（0 = 基期，1 = 跟踪期）和 x = 干预方式（0 = 安慰剂，1 = 现行安眠药）之间存在交互效应，但它假定结果变量的每个切点处的效应都相等。

该模型利用六个参数拟合 12 个边际 Logit。根据模型的最大似然拟合结果，比较单元格计数观察值和模型拟合值的检验统计量 $G^2 = 8.0$（df = 6）。关于参数的最大似然估计为 $\hat{\beta}_1 = 1.074$（SE = 0.162），$\hat{\beta}_2 = 0.046$（SE = 0.236），以及 $\hat{\beta}_3 = 0.662$（SE = 0.244）。结果表明，模型中存在交互效应。在研究基期，所估计的服用现行安眠药一组的入睡时

间低于任一给定水平的发生比等于服用安慰剂一组的相应发生比的 $\exp(0.046) = 1.04$ 倍;而在跟踪调查时,该效应为 $\exp(0.046 + 0.662) = 2.03$。换句话说,起初这两组具有相似的分布,但是在跟踪期,服用安眠药的一组总体上入睡所需的时间更短。

表 11.5 表 11.4 的样本边际分布

干预方式	调查时点	结果			
		<20	20~30	30~60	>60
现行药	基期	0.101	0.168	0.336	0.395
	跟踪期	0.336	0.412	0.160	0.092
安慰剂	基期	0.117	0.167	0.292	0.452
	跟踪期	0.258	0.242	0.292	0.208

在需要对分析结果进行较为简单的解释时,可以考虑报告样本的边际均值以及这些均值之间的差异。按照入睡时间的赋值{10,25,45,75},服用安眠药一组的基期均值为50.0,服用安慰剂一组的均值为50.3。分别比较两组在基期与跟踪期的均值可见,对于服用安眠药的一组,前后之差为22.2,而服用安慰剂一组的相应差异为13.0。两组均值的前后变动程度之间的差别等于9.2,其标准误为 SE = 3.0,这表明服用安眠药一组的变动明显更大。

11.2.4 控制结果变量基期取值后的对比

在类似表 11.4 的数据中,假定结果变量在所有干预组的基期边际分布是相同的。通过将研究对象进行随机分组可以保证,除抽样误差外,这一点是成立的。另外,假定以结果变量的基期取值为条件,各干预组在跟踪期的结果变量分布也相同。这样,跟踪期结果变量的边际分布也都相同。

然而,如果在基期时的边际分布不相同,即使跟踪期结果变量的条件分布相等,各组间跟踪期和基期的边际分布之差仍有可能不一样。在这种情况下,尽管边际模型仍然有用,但它可能并没有反映问题的全貌。因而,构建一个模型,在控制结果变量基期取值的基础上比较跟踪期的分布更有意义。

令 Y_2 表示跟踪期的结果变量,它对应的干预方式为 x,基期的结果变量为 y_1。在以下模型中:

$$\text{Logit}[P(Y_2 \leqslant j)] = \alpha_j + \beta_1 x + \beta_2 y_1, \tag{11.7}$$

β_1 比较在控制了基期观测值后不同干预方式对应的跟踪期结果变量的分布。这个模型与协方差分析(*analysis-of-covariance*)的原理相似,只不过这里的结果变量是定序变量而不是连续变量。这个累积 Logit 模型针对的是单个结果变量(Y_2)而不是多元结果变量(Y_1, Y_2)的边际分布。它是本章最后一节将要讨论的*转换模型*(*transitional model*)的一个例子。

11.2.5 边际 Logit 模型的最大似然拟合*

通过最大似然法拟合边际 Logit 模型非常不便。当对包括 I 个类别的结果变量进行了 T 次观测时,在预测变量的每个取值水平上似然函数包含 I^T 个多项分布联合概率,但模型所对应的却是 T 组多项分布的边际参数 $\{P(Y_t = k), k = 1, \cdots, I\}$。而且,这些多项分布的边际变量并不相互独立。

令 $\boldsymbol{\pi}$ 表示对应于预测变量所有取值水平的多项分布联合概率的完整集合。边际

Logit 模型具有如第 8.5.4 节所介绍的广义对数线性模型的形式：

$$\mathbf{C}\log(\mathbf{A}\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}。\tag{11.8}$$

在结果变量为二分变量的情况下，$\boldsymbol{\pi}$ 乘以矩阵 $\mathbf{A}$ 便得到预测变量的每个取值水平对应的 T 个边际概率 $\{P(Y_t = 1)\}$ 以及它们的补集。边际概率的对数乘以矩阵 $\mathbf{C}$ 形成了预测变量的每个取值水平对应的 T 个边际 Logit；在 $\mathbf{C}$ 的每一行中，与某个给定边际 Logit 的分子概率的对数相乘的位置对应的元素等于 1，在与分母概率的对数相乘的位置对应的元素等于 -1，其他元素等于 0。

例如，对于一个 2^T 表格，在不包括自变量的边际同质性模型中，$\boldsymbol{\beta}$ 为单个参数，在式 11.1 中由 α 表示。当 $T = 2$ 时，$\boldsymbol{\pi}$ 具有四个元素，相应的模型为

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}\log\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}\begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\alpha,$$

该模型设定 $\text{Logit}(\pi_{11} + \pi_{12}) = \text{Logit}[P(Y_1 = 1)]$ 和 $\text{Logit}(\pi_{11} + \pi_{21}) = \text{Logit}[P(Y_2 = 1)]$ 都等于 α。

边际 Logit 模型的似然函数 $l(\boldsymbol{\pi})$ 是在预测变量不同取值水平上的多项分布密度函数之间的乘积。最大似然拟合的一种方式将模型看作一组约束条件，并利用有关方法对约束限定下的函数进行最大化。在式 11.8 模型中，令 $\mathbf{U}$ 表示一个列满秩矩阵，使得由 $\mathbf{U}$ 的列所形成的空间是由 $\mathbf{X}$ 的列所形成的空间的正交补集。这时，$\mathbf{U}'\mathbf{X} = \mathbf{0}$，并且模型具有与下式等价的约束条件，

$$\mathbf{U}'\mathbf{C}\log(\mathbf{A}\boldsymbol{\pi}) = \mathbf{0}。$$

例如，按照上面式 11.8 的表述，在一个 2 × 2 表格的边际同质性模型中，$\mathbf{U}' = (1, -1)$。这时，将 $\mathbf{U}'$ 与 $\mathbf{C}\log(\mathbf{A}\boldsymbol{\pi})$ 相乘便可设定表格中行和列的边际 Logit 之差等于 0。

这种最大化似然函数的方法既考虑了模型本身的限定条件，也包括了模型的可识别性条件，后者限定在预测变量的每个取值水平上结果变量的概率之和等于 1。我们将模型的限定条件表达为 $\mathbf{U}'\mathbf{C}\log(\mathbf{A}\boldsymbol{\pi}) = \mathbf{0}$，将可识别性条件表达为 $\mathbf{f}(\boldsymbol{\pi}) = \mathbf{0}$。该方法引入了与这些约束条件相对应的拉格朗日算子，并通过 Newton-Raphson 算法求解拉格朗日似然方程（Aitchison and Silvey, 1958; Haber, 1985）。令 $\boldsymbol{\theta}$ 表示元素为 $\boldsymbol{\pi}$ 和拉格朗日算子为 $\boldsymbol{\lambda}$ 的向量。拉格朗日似然方程的形式为 $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$，其中

$$\mathbf{h}(\boldsymbol{\theta}) = (\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\lambda}) = (\mathbf{f}(\boldsymbol{\pi}), \partial\log[l(\boldsymbol{\pi})]/\partial\boldsymbol{\pi} + [\partial\mathbf{f}(\boldsymbol{\pi})/\partial\boldsymbol{\pi}]'\boldsymbol{\lambda})'$$

是一个向量，其元素与模型本身对边际 Logit 所做的设定以及对数似然函数的导数有关。

这时，由 Newton-Raphson 算法可得：

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left[\frac{\partial h(\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}}\right]^{-1} h(\boldsymbol{\theta}^{(t)}),\quad t = 1, \cdots。$$

这一运算非常复杂，因为对导数矩阵求逆所具有的维度比 $\boldsymbol{\pi}$ 中的元素数量还大。对此方法的一种改进是，利用重新参数化后的简化导数矩阵来进行渐近近似（Lang, 1996a; Lang and Agresti, 1994），此时只需要对一个对角矩阵和一个正定的对称矩阵求逆。

这种最大似然边际拟合方法可以通过专门的软件（附录 A 提到了 S-Plus 的一个命令）来应用。它对描述联合分布 $\boldsymbol{\pi}$ 的模型不需要做任何假定。因此，当边际模型成立时，不论联合分布具有什么形式的相依结构，关于式 11.8 中的 $\boldsymbol{\beta}$ 的最大似然估计是一致的。还有研究探讨过其他的拟合方法。Lang 和 Agresti（1994）同时拟合了关于 $\boldsymbol{\pi}$ 的边际模型

和非饱和对数线性模型。他们的完整模型可视为式 11.8 的一个特例,在拟合时使用了上文所介绍的带有拉格朗日算子的约束法。在一般情况下,边际模型参数和联合模型参数是正交的。如果边际模型成立,即使对联合分布所设定的模型是错误的,关于边际模型参数的最大似然拟合仍然是一致的。

Fitzmaurice 和 Laird(1993)给出了一种与此相关的最大似然法。$\boldsymbol{\pi}$ 和饱和对数线性模型的参数之间存在一一对应关系。他们进一步利用了对数线性模型的主效应和高阶参数与边际概率和相同的高阶对数线性参数之间的一一对应。这时,可以分别设定关于边际概率和关于高阶(条件)对数线性参数的模型。然后,对两组模型的参数求解最大化似然函数。同样,由于这两组参数是正交的,所以当边际模型成立时,关于它的参数的最大似然估计是一致的。这种*混合参数*(*mixed parameter*)法可以通过一些专门的软件(Kastner et al.,1997;另见附录 A)来应用。

另一种最大似然法利用 $\boldsymbol{\pi}$ 与描述边际分布、二元分布、三元分布等的参数之间的一一对应关系(如 Glonek and McCullagh,1995;Molenberghs and Lesaffre,1994)。这时,可以对各个分布项拟合多元 logistic 模型;为简便起见,可以考虑忽略掉一些高阶项。Glonek(1996)提出了一种关于此方法和 Fitzmaurice 和 Laird(1993)方法的混合方法。

11.3 边际模型分析:广义估计方程(GEE)法

最大似然拟合假定在预测变量的每个取值组合点,对包括 I 个类别的结果变量进行 T 次观测所形成的 I^T 个单元格概率服从多项分布。随着预测变量数量的增加,多项分布概率的数量会急剧增加。目前,当 T 很大或者存在很多预测变量时,尤其是当某些变量是连续变量时,前面介绍的所有最大似然方法都不现实。与应用多元正态分布假定的连续结果变量相比,多元分类结果变量的边际模型分析由于缺乏一个简单的多元分布来描述 T 个结果变量之间的关系而受到影响。例如,多元正态分布仅需要 $T+2$ 个参数,包括 T 个均值、一个共同的方差和相关系数,而在多项分布中却存在 I^T-1 个参数。

除最大似然拟合之外,另一种可以考虑的方法是对类似然法(第 4.7)的多元扩展。类似然法不需要假定 Y 服从某个特定分布,而仅需要指定分布的前两阶矩量;它将 Y 的均值与线性预测项相连结,并同时指定方差如何取决于均值。由该均值和方差指定一个指数族分布,满足相应分布的似然方程便构成了估计方程。对这些估计方程求解就得到了类似然法的估计值(Wedderburn,1974)。

11.3.1 广义估计方程法的基本思路

重复测量的多元结果变量为 $(Y_1,Y_2,\cdots,Y_T)$,其中不同对象所对应的 T 可能也不一样。与单变量的情况相同,类似然法设定一个关于 $\mu = E(Y)$ 的模型以及一个描述 $\text{var}(Y)$ 如何取决于 μ 的方差函数 $v(\mu)$。这时,可以对每个 Y_t 的边际分布进行模型分析。这种方法还需要对 $\{Y_t\}$ 之间的相关结构进行一个操作性猜测(working guess)。参数的估计值就是类似然方程的解,这些类似然方程被称为*广义估计方程*(*generalized estimating equations*)。因而,这种方法常称作 GEE 法。Liang 和 Zeger(1986)在拟合具有广义线性形式的边际模型时提出了该方法。他们的工作借鉴了经济计量学的有关文献(如 Gourieroux et al.,1984;Hansen,1982;White,1982)。在此,我们只介绍有关概念,具体细节将留到第 11.4 节。

GEE 法不需要假定某种特定的多元分布，它利用所假定的关于 $(Y_1, Y_2, \cdots, Y_T)$ 的协方差结构，设定相应的方差函数和成对相关模式。即使有关协方差结构的假定是错误的，关于模型参数的 GEE 估计仍然有效。估计的一致性（即估计值依概率收敛于参数的真值）取决于第一阶矩量，而与第二阶矩量无关。具体而言，假定模型选取的连结函数正确，且线性预测项真正描述 $E(Y_t)$ 如何取决于预测变量，其中 $t = 1, \cdots, T$，那么，关于模型参数的 GEE 估计具有一致性。

在现实应用中，几乎没有模型会是完全正确的。但是，上述结果仍是有意义的，因为它表明，相关结构的设定不会对所选模型的估计结果产生不良影响。通常来说，我们事先并不知道这个相关结构的任何信息，而且在数据分析中，这些相关关系一般被视为冗余参数。因此，为简便起见，在 GEE 法中可以简单地假定 $\{Y_t\}$ 之间两两独立。在这种简单假定下，尽管参数估计一般不会出现问题，但是对标准误的估计却不是这样。可以利用数据本身所展示的经验相依性对 GEE 法进行调整，以获得关于标准误的更适当的估计。利用数据所提供的有关相依结构的实际信息，调整在独立性假设下所得到的标准误，就可以得出更适当的(*稳健*(*robust*))标准误估计。

除了将 $\{Y_t\}$ 视为两两独立外，还可以在 GEE 法中对相关结构进行一个操作性猜测，然后再根据经验信息对标准误进行调整。*可交换的*(*exchangeable*)相关结构假定对所有的 s 和 t，$\text{corr}(Y_t, Y_s)$ 都相等。这比单纯的独立性假定更灵活，也更现实。比可交换结构更为现实的一种选择是无结构相关，即允许每对变量间都具有不同的相关模式。但是，当 T 很大时，由于这种无结构相关需要额外包括很多的参数，因而会导致一定的效率损失。

在理论上，选择适当的相关结构可以提高估计的效率。不过，Liang 和 Zegar(1986)指出，当实际的相关关系不是很强时，基于独立性相关结构假设的估计值具有非常好的效率。大家可以比较在不同相关结构假定下的拟合结果，检查模型对这些选择的灵敏度。根据我们的经验，当存在中等强度的相关关系时，无论选择哪种相关结构，得出的 GEE 估计值及其标准误都相似，因为经验的相依程度对调整初始的标准误具有很大的影响（如果不同假定下的结果相差很大，有必要对数据的相关结构进行进一步的研究）。除非事先预期到使用不同的相关结构假设会产生明显差异，我们建议使用可交换的相关结构。可交换的相关结构假定用一个额外的参数来反映可能的相依性。

对分类数据来说，GEE 法非常有吸引力，因为它在运算上比最大似然法更为简单。GEE 法的优点在于不需要设定一个具体的多元分布，而且估计的一致性不受相关结构设定错误的影响。然而，它也存在局限性。由于 GEE 法不指定完整的联合分布，所以在该方法中不具有一个似然函数。基于似然法的拟合检验、模型比较以及参数推断在此都不适用。相反，GEE 法中的统计推断使用沃尔德统计量，它是由估计值的渐近正态性以及所估计的协方差矩阵构建而来。但是，除非样本规模非常大，基于经验的标准误往往会低估真实的标准误（如：Firth，1993b）。与参数估计值一样，这些标准误也会显示出比参数估计更大的变动性（Kauermann and Carroll，2001）。Boos(1992)以及 Rotnitzky 和 Jewell(1990)利用类对数似然函数，提出了一种类似于关于预测变量效应的计分检验的检验，检验结果一般比沃尔德检验更可靠。由于 GEE 法缺乏似然函数，一些统计学家（如：Lindsey，1999）对该方法持否定态度；然而，其他统计学家并不认为这是一个问题，因为他们将 GEE 视为一种估计方法，而不是一个模型。

11.3.2　例子:精神抑郁的跟踪研究

对于表 11.2 比较治疗精神抑郁的两种方法的数据,我们在第 11.2.1 节利用最大似然法拟合了一个包含药品和时间交互项的 Logit 模型。利用 GEE 法来分析这些数据,无论选取哪种相关结构,都给出了相似的结果。在使用可交换结构时,GEE 法所估计的关于标准药物的斜率(以 Logit 为单位)为 $\hat{\beta}_3 = 0.48$ (SE =0.12),新药品对应的斜率则上升了 $\hat{\beta}_4 = 1.02$ (SE =0.19)。表 11.6 给出了使用独立性相关结构时的结果,该估计结果精确到小数点后两位数都与使用可交换结构的结果相一致。表 11.6 中的初始估计值及标准误是指当重复测量之间确实独立时相应的结果,相当于对 3 ×340 个独立观测值进行普通的 logistic 回归,而不是将数据视为对 340 个调查对象进行三次相依的重复测量。表中的经验标准误则是基于样本的相依性对独立性假定下的标准误进行调整后的结果。

表 11.6　利用 GEE 法对表 11.2 数据拟合 Logit 模型的输出结果

Initial Parameter Estimates			GEE Parameter Estimates Empirical std Error Estimates		
Parameter	Estimate	Std Error	Parameter	Estimate	Std Error
Intercept	-0.0280	0.1639	Intercept	-0.0280	0.1742
diagnose	-1.3139	0.1464	diagnose	-1.3139	0.1460
drug	-0.0596	0.2222	drug	-0.0596	0.2285
time	0.4824	0.1148	time	0.4824	0.1199
drug * time	1.0174	0.1888	drug * time	1.0174	0.1877

Working　Correlation　Matrix

	Col1	Col2	Col3
Row1	1.0000	0.0000	0.0000
Row2	0.0000	1.0000	0.0000
Row3	0.0000	0.0000	1.0000

译者注——Intercept:截距;diagnose:初期诊断结果;drug:药品;time:时点;drug * time:药品 * 时点

在使用可交换的相关结构时,所估计的三个结果变量中每对变量之间的共同相关系数为 -0.003,即同一对象的相继的观测值之间明显具有与独立观测值相似的特点。对于重复测量的数据来说,这种情况很少见。基于此,假定同一对象的三个观测值实际上来自于三个不同的对象(即,假定 1 020 个独立的观测值)来拟合模型可以得到相似的结果。

11.3.3　多项分布结果变量的 GEE 法:失眠症的例子

Liang 和 Zeger(1986)最初提出的 GEE 法是针对单变量边际分布分析的,比如二项分布和泊松分布模型。GEE 法也可以用来对多项分布结果变量的边际模型进行分析。Lipsitz 等(1994)对具有重复测量的定序结果变量的累积 Logit 模型提出了一种 GEE 法。按照该方法,可以对结果变量的每两个类别对应的重复观测值选择一种相关结构矩阵。在每个给定观测值处,每个多项分布结果对应的是 $(I-1)(I-1)$ 多项协方差矩阵。有关细节,可参见第 11.4.4 节。

我们以表 11.4 中关于失眠症的数据为例来对此加以说明。在第 11.2.3 节中,我们利用最大似然法拟合了边际模型

$$\text{Logit}[P(Y_t \leqslant j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 tx,$$

其中 Y_t = 使用治疗方式 x 的研究对象在时点 t 处的入睡时间。利用独立性相关结构，GEE 估计结果为 $\hat{\beta}_1 = 1.038$ (SE = 0.168)，$\hat{\beta}_2 = 0.034$ (SE = 0.238)，以及 $\hat{\beta}_3 = 0.708$ (SE = 0.244)。这些结果与最大似然估计结果相似，所得到的实质性结论也都相同。有相当的证据表明，治疗组入睡时间的分布比使用安慰剂一组下降得更快。

11.4 类似然法与 GEE 多元扩展：细节*

广义线性模型假定结果变量服从某种特定的分布。但有时候，我们并不清楚应该选择哪种分布。然而，结果变量的均值与方差之间往往都存在某种联系，比如对于计数数据有 $v(\mu_i) = \phi\mu_i$。这时，除最大似然估计之外，可以考虑使用类似然估计（第 4.7 节）。接下来，我们具体介绍一下类似然法，以及相应的关于多元结果变量边际模型分析的 GEE 扩展。

首先，我们考虑单一结果变量的模型，然后再讨论有关多元结果变量的边际模型。对于第 i 个研究对象，$i = 1,\cdots,n$，令 y_i 表示 Y 的结果，存在 $\mu_i = E(Y_i)$ 以及方差函数 $v(\mu_i)$，令 x_{ij} 表示第 j 个解释变量的取值。针对连结函数 g，线性预测项为 $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij} = \mathbf{x}_i'\boldsymbol{\beta}$。类似然（QL）参数估计值 $\hat{\boldsymbol{\beta}}$ 是类计分方程（quasi-score equations）

$$\mathbf{u}(\boldsymbol{\beta}) = \sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)' v(\mu_i)^{-1}(y_i - \mu_i) = \mathbf{0} \qquad (11.9)$$

的解，其中 $\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$。代入公式

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}x_{ij}$$

后，这些*估计方程*（*estimating equations*）与广义线性模型的似然方程（式 4.22）相同。但是，由于没有关于 $\{y_i\}$ 服从自然指数族分布的额外假定，它们并不是似然方程。在这种假定下，$v(\mu_i)$ 描述了该自然指数族分布的特征（Jϕrgensen，1987）。关于式 11.9 的另一个推导是，在利用已知方差 v_i 替代 $v(\mu_i)$ 后，它就是利用加权最小二乘法对 $\sum_i (y_i - \mu_i)^2 v_i^{-1}$ 求最小化的结果。

广义线性模型的似然方程（式 4.22）仅取决于 $\{y_i\}$ 的均值、方差以及连结函数 g，这些决定了 $\partial\mu_i/\partial\eta_i$。因此，Wedderburn（1974）建议，对所有的连结函数和方差函数，即使函数本身并不对应一个自然指数族分布，都使用式 4.22 作为估计方程。

11.4.1 类似然估计值的特性

在类似然（QL）法中，式 11.9 中的*类计分函数*（*quasi-score function*）$u_j(\boldsymbol{\beta})$ 被称为*无偏估计函数*（*unbiased estimating function*），意即，使得对所有 $\boldsymbol{\beta}$ 都满足 $E[h(\mathbf{Y};\boldsymbol{\beta})] = 0$ 的任意关于 $\mathbf{y}$ 和 $\boldsymbol{\beta}$ 的函数 $h(\mathbf{y};\boldsymbol{\beta})$。决定 $\hat{\boldsymbol{\beta}}$ 的方程式 11.9 被称为*估计方程*（*estimating equations*）。

在类似然法中，类计分函数就是所谓的*类对数似然函数*（*quasi-log likelihood*）的导数。类对数似然函数在严格意义上可能不是一个对数似然函数。不过，McCullagh（1983）表明，类似然估计值具有与最大似然估计值相似的特性。例如，类似然估计值 $\hat{\boldsymbol{\beta}}$ 渐近服从正态分布，其协方差矩阵近似为

$$V = \left[\sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)' [v(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right]^{-1}。 \tag{11.10}$$

它等价于广义线性模型的最大似然估计值之间的大样本协方差矩阵(其估计公式为式 4.28)。

一个至关重要的结论是,即使所设定的方差函数是错误的,只要连结函数和线性预测项是正确的,类似然估计值 $\hat{\boldsymbol{\beta}}$ 就是对 $\boldsymbol{\beta}$ 的一致估计(即 $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$)。换句话说,假定模型形式 $g(\mu_i) = \sum_j \beta_j x_{ij}$ 是正确的,即便真实的方差函数不是 $v(\mu_i)$,$\hat{\boldsymbol{\beta}}$ 的一致性仍然成立。现在,我们对此给出一个直观的解释。

当确实存在 $\mu_i = g^{-1}(\sum_j \beta_j x_{ij})$ 时,那么由式 11.9 可得,对所有 j 都有 $E[u_j(\boldsymbol{\beta})] = 0$。根据式 11.9,$\mathbf{u}(\boldsymbol{\beta})/n$ 是样本均值的向量,按照大数定理,它依概率收敛于其期望值 $\mathbf{0}$。类计分方程的解 $\hat{\boldsymbol{\beta}}$ 是这些样本均值的一个连续函数,因为 $\hat{\boldsymbol{\beta}}$ 是使相加之和正好等于 $\mathbf{0}$ 的 $\boldsymbol{\beta}$ 值,所以 $\hat{\boldsymbol{\beta}}$ 收敛于 $\boldsymbol{\beta}$。此外,一致性也可由无偏估计函数的一般结果推出(Liang and Zeger,1995)。

11.4.2 关于方差设定错误的 Sandwich 协方差调整

如果我们假定 $\text{var}(Y_i) = v(\mu_i)$ 但实际上 $\text{var}(Y_i) \neq v(\mu_i)$,那么类似然估计值 $\hat{\boldsymbol{\beta}}$ 的真正渐近协方差矩阵就不是由式 11.10 所给出的 $\mathbf{V}$。相反,它等于(Diggle et al.,2001;White,1982)

$$\mathbf{V}\left[\sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)' [v(\mu_i)]^{-1} \text{var}(Y_i) [v(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right] \mathbf{V}。 \tag{11.11}$$

尽管方差是一个标量(scalar),我们仍用这种形式表示矩阵,以便于下文对 GEE 多元扩展情况的讨论。如果 $\text{var}(Y_i) = v(\mu_i)$,式 11.11 就简化为 $\mathbf{V}$。在现实应用中,真实的方差函数是未知的。关于式 11.11 的一致性估计值是在它对应的样本形式中用 $\hat{\mu}_i$ 来替代 μ_i,并用 $(y_i - \hat{\mu}_i)^2$ 来替代 $\text{var}(Y_i)$ (Liang and Zeger,1986)。不论设定的方差函数 $v(\mu_i)$ 是否正确,由式 11.11 所估计的协方差矩阵都是有效的。这种关于协方差矩阵的估计被称为 *sandwich 估计值*(*sandwich estimator*),因为经验数据被模型设定的协方差矩阵像三明治一样包了起来。

总之,即便关于方差函数的设定是不正确的,我们仍然可以得到关于 $\boldsymbol{\beta}$ 的一致性估计,而且可以通过式 11.11 的 sandwich 调整来估计 $\hat{\boldsymbol{\beta}}$ 的渐近方差。但是,如果所选取的方差函数 $v(\mu_i)$ 与真实情况相差很大,这种方法会损失一部分效率。另外,为了使式 11.11的样本形式具有令人满意的结果,群组的数量 n 需要足够大,否则,它会倾向于低估真正的方差。当然,任何模型分析的过程都不可能保证所有的环节完全正确。正如所选取的方差函数只是对真实函数的近似一样(希望尽可能接近),所设定的均值也只是对真实值的近似。

11.4.3 GEE 法:技术细节

现在,我们考虑关于类似然法的广义估计方程(GEE)多元扩展。对于第 i 个对象,令 $\mathbf{y}_i = (y_{i1}, \cdots, y_{iT_i})'$ 以及 $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{iT_i})'$,其中 $\mu_{it} = E(Y_{it})$。结果变量的数量 T_i 在不同群组间可以变动。令 $\mathbf{x}_{it}$ 表示关于 y_{it} 的解释变量取值的一个 $p \times 1$ 向量。这种表述方式允许在重复测量中解释变量取值也不同的情况。对于连结函数 g,模型的线性预测项为 $\eta_{it} = g(\mu_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta}$。该模型所关注的是每个 t 点的边际分布,而不是联合分布。令 $\mathbf{X}_i$ 表

示关于第 i 个群组(或对象)的预测变量取值的一个 $T_i \times p$ 矩阵,其中第 t 行为 $\mathbf{x}'_{it}$。

我们假定 y_{it} 具有下列形式的概率密度函数:

$$f(y_{it};\theta_{it},\phi) = \exp\{[y_{it}\theta_{it} - b(\theta_{it})]/\phi + c(y_{it},\phi)\}。$$

当 ϕ 已知时,这是一个自然参数为 θ_{it} 的自然指数族分布。由第 4.4.1 节可得,

$$\mu_{it} = E(Y_{it}) = b'(\theta_{it}),\quad v(\mu_{it}) = \mathrm{var}(Y_{it}) = b''(\theta_{it})\phi。$$

同时,GEE 法假定 $\mathbf{Y}_i$ 具有操作性相关结构 $\mathbf{R}(\boldsymbol{\alpha})$,该相关结构取决于参数 $\boldsymbol{\alpha}$。在可交换的相关结构下,$\mathbf{Y}_i$ 中的每两个变量都有 $\mathrm{corr}(Y_{it},Y_{is}) = \alpha$。令 $\mathbf{b}_i(\boldsymbol{\theta}) = (b(\theta_{i1}),\cdots,b(\theta_{iT_i}))$,并令 $\mathbf{B}_i$ 表示主对角线元素为 $\mathbf{b}''_i(\boldsymbol{\theta})$ 的对角矩阵。这时,关于 $\mathbf{Y}_i$ 的操作性协方差矩阵为

$$\mathbf{V}_i = \mathbf{B}_i^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{B}_i^{1/2}\phi。\tag{11.12}$$

注意,如果 $\mathbf{R}$ 是 $\mathbf{Y}_i$ 的真实协方差矩阵,那么 $\mathbf{V}_i = \mathrm{cov}(\mathbf{Y}_i)$。

现在令 $\boldsymbol{\Delta}_i$ 表示主对角线元素为 $\partial\theta_{it}/\partial\eta_{it}$ 的对角矩阵,$t = 1,\cdots,T_i$(在典型连结的情况下,它等于恒等矩阵)。令 $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta} = \mathbf{B}_i\Delta_i\mathbf{X}_i$ 表示一个 $T_i \times p$ 矩阵,其主要元素是表示为 $(\partial\mu_{it}/\partial\theta_{it})(\partial\theta_{it}/\partial\eta_{it})(\partial\eta_{it}/\partial\beta_j)$ 形式的 $\partial\mu_{it}/\partial\beta_j$。根据式 11.9,单变量广义线性模型的类似然估计方程具有以下形式:

$$\sum_i\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right)'v(\mu_i)^{-1}[y_i - \mu_i(\boldsymbol{\beta})] = \mathbf{0},$$

其中 $\mu_i = \mu_i(\boldsymbol{\beta}) = g^{-1}(\mathbf{x}'_i\boldsymbol{\beta})$。在多元情况下,相应的一组*广义估计方程*(*generalized estimating equations*)为

$$\sum_{i=1}^{n}\mathbf{D}'_i\mathbf{V}_i^{-1}[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0}。$$

GEE 估计值 $\hat{\boldsymbol{\beta}}$ 就是上述方程组的解。

设定 $\mathbf{R}(\boldsymbol{\alpha}) = \mathbf{I}$ 的简单方法(naive approach)将每对结果变量都视为相互独立的。在此情况下,式 11.12 简化为 $\mathbf{V}_i = \mathbf{B}_i\phi$,并且相应的广义估计方程简化为

$$\begin{aligned}\sum_i\mathbf{D}'_i\mathbf{V}_i^{-1}[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] &= \sum_i\mathbf{X}'_i\Delta_i\mathbf{B}_i\mathbf{V}_i^{-1}[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})]\\ &= (1/\phi)\sum_i\mathbf{X}'_i\boldsymbol{\Delta}_i[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0},\end{aligned}$$

即 $\sum_i\mathbf{X}'_i\boldsymbol{\Delta}_i[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0}$。这时,它的解 $\hat{\boldsymbol{\beta}}$ 与将 $(y_{i1},\cdots,y_{iT_i})$ 视为独立观测值并根据所选定的连结函数和方差函数拟合普通广义线性模型的结果相同。

一般来说,我们会选择一个允许相依性的操作性相关矩阵,如可交换的相关结构。在分析时间序列数据时,自回归结构也是一个常见的选择,它满足 $\mathrm{corr}(Y_{it},Y_{is}) = \alpha^{|t-s|}$,即认为在时间上相距较远的观测值之间的相关程度也较弱。Liang 和 Zeger(1986)建议,通过利用求解关于 $\boldsymbol{\beta}$ 的广义估计方程组的修正 Fisher 计分算法(给定对 $\boldsymbol{\alpha}$ 和 ϕ 的当前估计)与利用残差对 $\boldsymbol{\alpha}$ 和 ϕ 进行的矩估计(基于对 $\boldsymbol{\beta}$ 的当前估计)之间的迭代,可以计算 GEE 估计值。他们也给出了在各种相关结构下关于 $\mathbf{R}(\boldsymbol{\alpha})$ 的估计。还有其他算法同时求解关于 $\boldsymbol{\beta}$ 和关联参数的估计方程组(如:Liang et al.,1992;另见注解 11.8)。GEE 算法不需要收敛,通常一次迭代就可以得到不错的结果(Lipsitz et al.,1991)。

Liang 和 Zeger(1986)表明,随着群组(cluster)数量 n 的增加,GEE 估计值具有渐近正态性和一致性。在一定的规则性条件下,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0},\mathbf{V}_G)。$$

这里,通过对式 11.11 进行扩展,$\mathbf{V}_G = \lim_{n\to\infty}\mathbf{V}_{G,n}$,其中

$$\mathbf{V}_{G,n} = n\left[\sum_i \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right]^{-1}\left[\sum_i \mathbf{D}_i'\mathbf{V}_i^{-1}\mathrm{cov}(\mathbf{Y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i\right]\left[\sum_i \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right]^{-1}。$$

关于 $\hat{\boldsymbol{\beta}}$ 的协方差矩阵的估计值 $\hat{\mathbf{V}}_{G,n}/n$ 用 $\hat{\boldsymbol{\beta}}$ 替代 $\boldsymbol{\beta}$、$\hat{\phi}$ 替代 ϕ、$\hat{\alpha}$ 替代 α 以及 $[\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})][\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})]'$ 替代 $\mathrm{cov}(\mathbf{Y}_i)$。Sandwich 估计值的目的在于,当操作性猜测与实际的相关结构相去甚远时,利用数据中关于相关关系的经验信息来调整标准误。

当操作性相关结构与现实情况相同并且 $\mathrm{cov}(\mathbf{Y}_i) = \mathbf{V}_i$ 时,渐近协方差矩阵 $\hat{\mathbf{V}}_{G,n}/n$ 简化为 $(\sum_i \mathbf{D}'_i\mathbf{V}_i^{-1}\mathbf{D}_i)^{-1}$。这是当我们对相关结构的猜测具有完全自信时所对应的协方差矩阵。

在二分数据的情况下,相关系数可能并不是表示组内关联的最优方式。这时,由于 $E(Y_{it}Y_{is}) = P(Y_{it} = 1, Y_{is} = 1)$ 的取值范围取决于 $P(Y_{it} = 1)$ 和 $P(Y_{is} = 1)$,边际概率限定了相关系数的可能取值。一种办法是使用发生比之比,比如在模型分析中将同一群组内每对观测值的对数发生比之比视为可交换的。这种方法的优点在于,它将关联参数与有关均值的模型明确区分开来,参见:Fitzmaurice 等(1993)以及 Lipsitz 等(1991)。Carey 等(1993)提出了一种迭代*交替 logistic 回归*(*alternating logistic regression*)算法。在这种算法中,关于均值的模型回归参数的 GEE 拟合步骤与关于发生比之比的关联模型的拟合步骤交替进行。当关联结构本身也是分析所关注的重点时,这种算法非常有用。

11.4.4 GEE 法:多项分布结果变量

现在,我们简要介绍一下 Lipsitz 等(1994)提出的拟合多项分布结果变量的边际模型的 GEE 法。例如,该方法可以应用于累积 Logit 模型。如果第 i 个群组内的第 t 个观测值对应的结果为 $j(j = 1, \cdots, I-1)$,记 $y_{it}(j) = 1$。令 $\mathbf{y}_i$ 表示关于第 i 个群组的 $T_i(I-1)$ 个二分指示变量。这时,我们可以选择一个关于 $\mathbf{y}_i$ 的 $[T_i(I-1)] \times [T_i(I-1)]$ 操作性协方差矩阵 $\mathbf{V}_i$,设定每对结果变量类别 (j,k) 以及每对 (t,s) 之间的 $\mathrm{corr}(Y_{it}(j), Y_{is}(k))$ 的模式。$\mathbf{V}_i$ 中关于 $(y_{it}(1), \cdots, y_{it}(I-1))$ 的 $(I-1)(I-1)$ 子阵是一个多项分布协方差矩阵,其主对角线元素为 $v_{it}(j) = P(Y_{it}(j) = 1)[1 - P(Y_{it}(j) = 1)]$,主对角线以外的元素为 $-P(Y_{it}(j) = 1)P(Y_{it}(k) = 1)$。$\mathbf{V}_i$ 的其他元素包含了 $\mathrm{cov}(Y_{it}(j), Y_{is}(k))$。例如,一种可能的选择是可交换的结构,其中对所有 t 和 s, $\mathrm{corr}(Y_{it}(j), Y_{is}(k)) = \rho_{jk}$。

按照这一方法,关于 $\boldsymbol{\beta}$ 的广义估计方程同样具有如下形式:

$$\mathbf{u}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

其中 $\boldsymbol{\mu}_i$ 是关于 $\mathbf{y}_i$ 的概率向量,$\mathbf{D}_i' = \partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta}$,并且参数等于它们当前的估计值。Lipsitz 等建议使用 Fisher 计分法来求解这些方程,并通过矩估计方法在每次迭代中更新对 $\{\rho_{jk}\}$ 的估计。对 $\hat{\boldsymbol{\beta}}$ 进行经验调整的 sandwich 协方差矩阵仍为

$$\left[\sum_{i=1}^{n} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right]^{-1}\left[\sum_{i=1}^{n} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathrm{cov}(\mathbf{Y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i\right]\left[\sum_{i=1}^{n} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right]^{-1}。$$

这个矩阵可以通过从模型拟合结果中代入 $\hat{\boldsymbol{\mu}}_i$ 并利用 $\mathbf{y}_i$ 的经验协方差矩阵替代 $\mathrm{cov}(\mathbf{Y}_i)$ 来估计。

11.4.5 缺失数据的处理

遗憾的是,在重复测量的研究中,往往会出现在某个群组内至少有一个结果变量的取值缺失的情况。例如,在一项跟踪调查中,有些对象可能会在研究结束之前退出。当存在缺失数据时,只分析所观察到的数据而忽视缺失数据的存在可能导致有偏的估计

结果。

GEE 法的一个优点在于,不同群组可以具有不同数量的观测值。在数据文件中每个观测值都对应着单独的一行,对于跟踪研究而言,运算过程使用的是那些每个研究对象参与观测的次数。然而,除非我们对数据的缺失机理做一定的假设,否则 GEE 估计结果仍然可能是有偏的。

令 $\mathbf{Y}^{(o)}$ 表示所观测到的结果变量,$\mathbf{Y}^{(m)}$ 表示缺失的结果变量,$\mathbf{Y}$ 表示所有结果变量。令 M 表示关于数据缺失情况的指示变量,当一个观测值缺失时,它等于 1,否则它等于 0。如果 M 在统计上独立于 $\mathbf{Y}$,Little 和 Rubin(1987)称之为*完全随机缺失*(*missing completely at random*),也即,一个观测值缺失的概率独立于该观测值的取值,尽管它可能取决于某些解释变量。进一步放松条件,如果 $(M\mid \mathbf{Y})$ 的分布等于 $(M\mid \mathbf{Y}^{(o)})$ 的分布,它们将之称为*随机缺失*(*missing at random*),即是说,缺失与否仅取决于 $\mathbf{Y}^{(o)}$,而与那些缺失值无关。

当数据缺失满足上述两种情况时,利用似然法进行分析不必要在模型中考虑数据的缺失机理。在分析中只使用 $\mathbf{Y}^{(o)}$ 的信息不会导致系统性偏差。当估计方程可以按照结果变量的概率进行加权时,以上结论对 GEE 法也成立(Robins et al.,1995)。其他情况下,对于诸如 GEE 法的不基于似然函数的分析方法,只有当数据满足完全随机缺失时,才可以忽略缺失的机理。Kenward 等(1994)表明,当数据不满足完全随机缺失时,GEE 估计会出现问题。

通常情况下,数据缺失的机理都与缺失值有关。例如,在一项测量疼痛感的跟踪研究中,当疼痛感超过某个临界值时,该研究对象就可能会选择退出。这时,有必要对 $\mathbf{Y}$ 和 M 的联合分布建立模型,进行更复杂的分析(Little,1998)。令 $f(\cdot)$ 表示一个通用的概率密度函数,它同时取决于解释变量 $\mathbf{x}$ 和模型参数。*选择模型*(*selection models*)将 $\mathbf{Y}$ 和 M 的联合分布分解为

$$f(\mathbf{y},M;\mathbf{x},\boldsymbol{\beta},\boldsymbol{\psi}) = f(\mathbf{y};\mathbf{x},\boldsymbol{\beta})f(M\mid \mathbf{y};\mathbf{x},\boldsymbol{\psi}),$$

其中 $f(\mathbf{y};\mathbf{x},\boldsymbol{\beta})$ 是不存在缺失值时的模型,$f(M\mid \mathbf{y};\mathbf{x},\boldsymbol{\psi})$ 是关于数据缺失机理的模型。*混合模式模型*(*pattern mixture models*)使用的是另一种因式分解,

$$f(\mathbf{y},M;\mathbf{x},\boldsymbol{\theta},\boldsymbol{\Phi}) = f(\mathbf{y}\mid M,\mathbf{x},\boldsymbol{\Phi})f(M;\mathbf{x},\boldsymbol{\theta}),$$

其中 $\mathbf{Y}$ 的分布以数据缺失的模式为条件。当 M 独立于 $\mathbf{Y}$ 时,这两种模型是等价的,也即存在 $\boldsymbol{\beta} = \boldsymbol{\Phi}$ 和 $\boldsymbol{\psi} = \boldsymbol{\theta}$。关于每种模型的优缺点以及缺失数据的各种处理方法的详细讨论,参见:Little(1988),以及注解 11.9 给出的文献。将缺失模式加入模型以检查缺失过程是否与结果变量相关或者是否与解释变量存在交互效应的例子,参见:Stokes 等(2000,p.524)。

在存在大量缺失数据的情况下,数据分析应当格外谨慎。一般来说,我们对数据缺失的机理知之甚少,而且相应的假定往往无法验证。这时,推断结果可能并不稳定,因而有必要通过灵敏度分析来检查结果是否会因数据缺失机理的不同假定而发生变动。当缺乏对缺失机理的模型分析时,我们至少应当比较使用所有群组的所有可知观测值与只使用不存在缺失的群组的观测值进行分析所得的结果。如果这两个结果存在很大差别,除非我们能够进一步分析缺失的机理,否则任何结论都只能是尝试性的。

11.5 马尔科夫链:转换模型

当 Y_t 表示在时点 t 对应的结果变量时,$t = 0,1,2,\cdots$,相应的随机变量族 $(Y_0,Y_1,Y_2,\cdots)$ 构成一个*随机过程*(*stochastic process*)。这个过程的*状态空间*(*state space*)是关于

Y_t 的一组可能取值,其中,Y_0 的值表示*初始状态*(*initial state*)。在状态空间为分类的且观测值发生的时间是离散的情况下,$\{Y_t\}$ 具有*离散状态空间*(*discrete state space*)和*离散时间*(*discrete time*)。

11.5.1 转换模型

一般情况下,研究关注的焦点主要是 Y_t 对以往各期的观测值($y_0, y_1, \cdots, y_{t-1}$)以及其他解释变量的相依性。这种类型的模型被称为*转换模型*(*transitional models*)。令 $f(y_0, \cdots, y_T)$ 表示 $(Y_0, \cdots, Y_T)$ 的联合概率密度函数(在这里,暂时不考虑解释变量)。转换模型可以进行以下因式分解:

$$f(y_0, \cdots, y_T) = f(y_0)f(y_1 \mid y_0)f(y_2 \mid y_0, y_1)\cdots f(y_T \mid y_0, y_1, \cdots, y_{T-1})。$$

与本章中其他几节所介绍的边际模型不同,转换模型分析以结果变量在以往各期的取值为条件。

在本节中,我们介绍离散时间*马尔科夫链*(*Markov chains*),这是一种具有离散状态空间的简单随机过程。许多转换模型至少部分具有马尔科夫结构。

11.5.2 一阶马尔科夫链

马尔科夫链(*Markov chain*)是这样一个随机过程,对于所有 t,给定 $Y_0, \cdots, Y_t$ 时 Y_{t+1} 的条件分布与只给定 Y_t 时的条件分布相同。换句话说,给定 Y_t,Y_{t+1} 条件独立于 $Y_0, \cdots, Y_{t-1}$。对于一个马尔科夫链,知道其当前状态后,其过去状态的信息对我们预测它的未来状态没有任何帮助。在马尔科夫链中,

$$f(y_0, \cdots, y_T) = f(y_0)f(y_1 \mid y_0)f(y_2 \mid y_1)\cdots f(y_T \mid y_{T-1})。 \quad (11.13)$$

如果对于所有 t,给定 $Y_0, \cdots, Y_t$ 时 Y_{t+1} 的条件分布与只给定 $(Y_t, \cdots, Y_{t-k+1})$ 时的条件分布相同,这个随机过程被称为 *k 阶马尔科夫链*(*k th-order Markov chain*)。给定前 k 次的状态,该链的未来取值独立于在那些 k 次之前的状态。在这里,我们的讨论主要集中于如式 11.13 所示的普通马尔科夫链,即一阶马尔科夫链($k = 1$)。

用 $\pi_{j|i}(t)$ 表示条件概率 $P(Y_t = j \mid Y_{t-1} = i)$,则这些 $\{\pi_{j|i}(t)\}$ 被称为*转换概率*(*transition probabilities*),它们满足 $\sum_j \pi_{j|i}(t) = 1$。$\{\pi_{j|i}(t), i = 1, \cdots, I, j = 1, \cdots, I\}$ 是一个 $I \times I$ 阶的*转换概率矩阵*(*transition probability matrix*)。它也被称为*一步*(*one-step*)矩阵,以区别于从时点 $t - k$ 到时点 t 的 k 步转换概率矩阵。

根据式 11.13,马尔科夫链的联合分布只取决于一步转换概率及其初始状态的边际分布。由此同时可以推出,它的联合分布满足对数线性模型

$$(Y_0Y_1, Y_1Y_2, \cdots, Y_{T-1}Y_T)。$$

对于随机过程实际结果的一个样本,列联表显示了可能序列的计数。对该对数线性模型的拟合结果进行检验,可以用来检查相应过程是否满足马尔科夫特性。

分类数据分析的标准方法可用于进行关于马尔科夫链的统计推断。例如,考虑有关转换概率的最大似然估计。令 $n_{ij}(t)$ 表示由时点 $t - 1$ 到时点 t 从状态 i 转换到状态 j 的数量。对于固定的 t,$\{n_{ij}(t)\}$ 是一个由 I^{T+1} 维列联表中第 $t - 1$ 和第 t 个维度所组成的二维边际表格。对于在时点 $t - 1$ 处状态为类别 i 的 $n_{i+}(t)$ 个对象,假定 $\{n_{ij}(t), j = 1, \cdots, I\}$ 服从参数为 $\{\pi_{j|i}(t)\}$ 的多项分布。令 $\{n_{i0}\}$ 表示初始的计数,假定这些初始计数也服从多项分布,其参数为 $\{\pi_{i0}\}$。如果各对象的行为相互独立,由式 11.13 可得,似然函数与

$$\left(\prod_{i=1}^{I} \pi_{i0}^{n_{i0}}\right)\left\{\prod_{t=1}^{T}\prod_{i=1}^{I}\left[\prod_{j=1}^{I} \pi_{j|i}(t)^{n_{ij}(t)}\right]\right\} \quad (11.14)$$

成比例。转换概率是 IT 个相互独立的多项分布的参数。根据 Anderson 和 Goodman (1957),对其的最大似然估计等于

$$\hat{\pi}_{j|i}(t) = n_{ij}(t)/n_{i+}(t)。$$

11.5.3 例子:呼吸道疾病

表 11.7 取自一项由哈佛大学所进行的跟踪研究,用以考察大气污染对儿童呼吸道疾病的影响。这些儿童在 9 岁至 12 岁之间每年接受体检,并按照是否患有哮喘作了划分。

由 Y_t 表示在年龄 t 时的二分结果变量(哮喘,无哮喘),$t = 9,10,11,12$。对数线性模型 $(Y_9Y_{10}, Y_{10}Y_{11}, Y_{11}Y_{12})$ 代表一阶马尔科夫链。它的拟合结果很差,$G^2 = 122.9$ (df = 8)。因而,给定在时点 t 的状态,时点 $t+1$ 的取值仍然取决于时点 t 之前的状态。模型 $(Y_9Y_{10}Y_{11}, Y_{10}Y_{11}Y_{12})$ 代表二阶马尔科夫链,即给定在 10 岁和 11 岁的状态后,9 岁和 12 岁的状态之间满足条件独立。这个模型的拟合结果也不好,$G^2 = 23.9$ (df = 4)。究其原因,这可能部分反映了对象间的异质性的影响,因为在上述分析中并没有考虑可能的协变量,比如父母的吸烟行为。

允许每两个年龄间都存在关联的对数线性模型 $(Y_9Y_{10}, Y_9Y_{11}, Y_9Y_{12}, Y_{10}Y_{11}, Y_{10}Y_{12}, Y_{11}Y_{12})$ 拟合得很好,$G^2 = 1.5$ (df = 5)。表 11.8 给出了有关成对的条件对数发生比之比的最大似然估计。相隔 1 岁的每对年龄之间的关联看上去相似,而相隔超过 1 岁的两个年龄之间的关联则较弱。满足

$$\lambda_{ij}^{Y_9Y_{10}} = \lambda_{ij}^{Y_{10}Y_{11}} = \lambda_{ij}^{Y_{11}Y_{12}} \quad 和 \quad \lambda_{ij}^{Y_9Y_{11}} = \lambda_{ij}^{Y_9Y_{12}} = \lambda_{ij}^{Y_{10}Y_{12}}$$

的更简单模型也拟合得很好,$G^2 = 2.3$ (df = 9)。对于前者,所估计的对数发生比之比为 1.75,后者则等于 1.04。

表 11.7 在四个年龄进行呼吸测试的结果[a]

Y_9	Y_{10}	Y_{11}	Y_{12}	计数	Y_9	Y_{10}	Y_{11}	Y_{12}	计数
1	1	1	1	94	2	1	1	1	19
1	1	1	2	30	2	1	1	2	15
1	1	2	1	15	2	1	2	1	10
1	1	2	2	28	2	1	2	2	44
1	2	1	1	14	2	2	1	1	17
1	2	1	2	9	2	2	1	2	42
1	2	2	1	12	2	2	2	1	35
1	2	2	2	63	2	2	2	2	572

a 1,哮喘;2,无哮喘。

来源:Ware et al. (1988)。

表 11.8 对表 11.7 数据所估计的条件对数发生比之比

关 联	估计值	较简单的结构
Y_9Y_{10}	1.81	1.75
$Y_{10}Y_{11}$	1.65	1.75
$Y_{11}Y_{12}$	1.85	1.75
Y_9Y_{11}	0.95	1.04
Y_9Y_{12}	1.05	1.04
$Y_{10}Y_{12}$	1.07	1.04

11.5.4 包括解释变量的转换模型

一般情况下,转换模型也包括解释变量 $\mathbf{x}$。这时,关于 T 个序列结果变量的联合密度函数为:

$$\begin{aligned}&f(y_1,\cdots,y_T;\mathbf{x})\\&=f(y_1;\mathbf{x})f(y_2\mid y_1;\mathbf{x})f(y_3\mid y_1,y_2;\mathbf{x})\cdots f(y_T\mid y_1,y_2,\cdots,y_{T-1};\mathbf{x})。\end{aligned}$$

例如,在 y 为二分变量的情况下,大家可以对上述因式分解中的每一项都设定一个 logistic 回归模型,

$$\begin{aligned}&f(y_t\mid y_1,\cdots,y_{t-1};\mathbf{x}_t)\\&=\frac{\exp[y_t(\alpha+\beta_1y_1+\cdots+\beta_{t-1}y_{t-1}+\boldsymbol{\beta}'\mathbf{x}_t)]}{1+\exp(\alpha+\beta_1y_1+\cdots+\beta_{t-1}y_{t-1}+\boldsymbol{\beta}'\mathbf{x}_t)},y_t=0,1。\end{aligned}$$

这里,针对不同项,预测变量 $\mathbf{x}$ 可以取不同的值。这个模型将前面所发生的结果变量也作为解释变量,称为*递归 logistic 模型*(*regressive logistic model*)(Bonney,1987)。

对 $\hat{\boldsymbol{\beta}}$ 的解释及其强度取决于在模型中有多少个前期观测值。在以前期的结果变量为条件后,组内效应(within-cluster effects)可能会急剧下降。这是转换模型与边际模型之间的一个重要区别,对边际模型的解释并不取决于所设定的相依结构。在一阶马尔科夫结构的特殊情况下,关于 y_t 的模型中,$\{y_1,\cdots,y_{t-2}\}$ 的相应系数等于0(如:Azzalini,1994;Bonney,1987)。有时,允许 $\mathbf{x}_t$ 和 y_{t-1} 具有对 y_t 的交互效应可能会很有意义。

对于一个给定的研究对象,相应的条件密度函数的乘积决定了该对象对似然方程的贡献(第一项的边际分布的贡献往往被忽略)。也就是说,给定预测变量的取值,模型将同一对象所发生的重复转换视为独立的。因而,可以利用有关普通广义线性模型的软件来拟合该模型,将每次转换分别视为一个单独的观测值(Bonney,1986)。

11.5.5 儿童呼吸道疾病与母亲的吸烟行为

表 11.9 同样取自哈佛大学对大气污染与健康关系的研究。在 7 岁至 10 岁间,孩子们每年接受关于呼吸道疾病的检查,其中一个预测变量是在研究开始时母亲的吸烟行为,这里 $s=1$ 表示母亲有吸烟习惯,否则 $s=0$。令 y_t 表示在年龄 t 时的结果变量($t=7,8,9,10$)。我们考虑递归 logistic 模型

$$\text{Logit}[P(Y_t=1)]=\alpha+\beta_1s+\beta_2t+\beta_3y_{t-1},\quad t=8,9,10。$$

表 11.9 按年龄和母亲吸烟行为划分的儿童呼吸道疾病情况

儿童呼吸道疾病			母亲不吸烟		母亲吸烟	
			10 岁		10 岁	
7 岁	8 岁	9 岁	否	是	否	是
否	否	否	237	10	118	6
		是	15	4	8	2
	是	否	16	2	11	1
		是	7	3	6	4
是	否	否	24	3	7	3
		是	3	2	3	1
	是	否	6	2	4	2
		是	5	11	4	7

来源:数据由 James Ware 友情提供。

每个对象都对模型拟合贡献了三个观测值。该数据包括 12 个二项分布，对应着 (s,t,y_{t-1}) 的 $2\times3\times2$ 种组合。例如，对于组合 (0,8,0)，有 $237+10+15+4=266$ 个对象存在 $y_8=0$，有 $16+2+7+3=28$ 个对象存在 $y_8=1$。关于该模型的最大似然拟合结果为

$$\text{Logit}[\hat{P}(Y_t=1)]=-0.293+0.296s-0.243t+2.211y_{t-1},$$

参数估计的标准误分别为 (0.846,0.156,0.095,0.158)。结果并不令人意外，前期的观测值具有很强的影响。给定前期状况和孩子的年龄，有轻微证据显示母亲的吸烟行为具有正效应：有关 $H_0: \beta_1=0$ 的似然比统计量为 3.55(df = 1, $P=0.06$)。此外，模型本身没有表现出任何拟合不充分的证据 ($G^2=3.1$, df = 8)。

注 解

第 11.1 节：边际分布的比较：多元结果变量的情况

11.1 Darroch(1981)系统回顾了边际同质性检验的统计量之间的关系以及它们与广义 CMH 分析的联系。另见：Mantel 和 Byar(1978)、White 等(1982)。在广义对数线性模型的框架下，Croon 等(2000)分析了关于跟踪数据的多种假设。

第 11.2 节：边际模型分析：最大似然法

11.2 有关利用最大似然法拟合边际模型的其他研究，参见：Bergsma 和 Rudas(2002)、Ekholm 等(2000)、Fitzmaurice 等(1993)、Lang 等(1999)。

第 11.3 节：边际模型分析：广义估计方程(GEE)法

11.3 Liang 等(1992)讨论了关于分类结果变量(主要是二分变量)的 GEE 法。关于多项分布结果变量的情况，参见：Heagerty 和 Zeger(1996)、Lipsitz 等(1994)，Miller 等(1993)，以及 Agresti 和 Natarajan(2001)所提到的文献。在关于定序结果变量的更一般化的模型中，允许离散参数也随着协变量的变动而变动(Toledano and Gatsonis,1996)。

11.4 LaVange 等(2001)利用 GEE 法来对调查和临床试验中的整群抽样进行调整。Boos(1992)探讨了引入经验方差估计的广义计分检验，并给出了在二分变量回归中有关趋势检验以及拟合不足检验的例子。

11.5 Koch 等(1977)通过加权最小二乘法(WLS)拟合了关于表 11.2 的边际模型。第 15.1 节介绍了对分类变量进行模型分析的 WLS 法。这种方法存在严重缺陷(例如，协变量必须是分类变量而且边际表格不能存在稀疏数据的问题)，但由它可以自然地推出 GEE 法。

第 11.4 节：类似然法与 GEE 多元扩展：细节

11.6 Firth(1993b)对类似然法进行了有益的综述。McCullagh(1983)证明，当对均值和方差函数的设定正确时，在满足关于 $\{y_i\}$ 局部线性的估计值中，类似然估计值具有渐近有效性。他的结果是对高斯-马尔科夫定理(Gauss-Markov theorem)的扩展，尽管这种扩展采用的是一种渐近的而不是精确的方式。有关无偏估计函数及其与渐近一致性、有效性的联系，参见：Heyde(1997)、Liang 和 Zeger(1995)。在 1960 年 Godambe 证明，最大似然估计值是一个无偏估计函数的最优解。当类似然估计值与最大似然估计值不同时，Cox(1983)以及 Firth(1987)指出，如果对自然指数族的偏离不是很严重——比如出现了一定程度的过度离散，类似然估计值仍具有很好的有效性。

11.7 在某些情况下，广义估计方程与似然方程相同，因而 GEE 估计也与最大似然估计相同。这样的例子包括多元正态数据或者操作性协方差矩阵设定正确的二分数

据(Fitzmaurice et al.,1993)。在各种模型构建的过程中,出现了许多关于模型设定错误的后果的研究。有关的一般性理论,参见:Gourieroux 等(1984)、Hansen(1982)、Liang 和 Zeger(1995)、White(1982)。

11.8　一种叫作 GEE2 的分析加入了关于相关结构的估计方程(Prentice and Zhao,1991)。这种做法可能会提高效力。它的缺点在于,与普通 GEE 不同,如果模型中关于相关结构这部分的设定有误,$\hat{\beta}$ 不再具有一致性。Qu 等(2000)展示了如何通过将操作性相关矩阵表示为有关基矩阵的线性组合来提高效力。

11.9　有关处理缺失数据的研究综述,参见:Little(1988)、Little 和 Rubin(1987,Chap.9)、Schafer(1997)、Verbeke 和 Molenberghs(2000)。另见:Baker 和 Laird(1988)、Fay(1986)、Fitzmaurice 等(1994)、Foster 和 Smith(1998)、Fuchs(1982)、Molenberghs 和 Goetghebeur(1997)、Molenberghs 等(1997)、Park 和 Brown(1994)、Stokes 等(2000)。

第 11.5 节:马尔科夫链:转换模型

11.10　有关马尔科夫链的统计推断,参见:Anderson(1980,Sec.7.7)、Anderson 和 Goodman(1957)、Billingsley(1961)、Bishop 等(1975,Chap.7)、Kalbfleisch 和 Lawless(1985)。有关条件相依结构的其他分析,参见:Conaway(1989)、Stiratelli 等(1984)、Ware 等(1988)。

习　题

应用部分

11.1　参见表 8.3。将该表视为三个一组的配对数据,分别构建关于学生每种行为的边际分布。求学生服用大麻、饮酒以及吸烟的样本比例。对边际同质性假设进行检验,并解释结果。

11.2　参见表 9.1。拟合一个关于种族、性别和行为类型(服用大麻、饮酒、吸烟)对学生是否具有该行为的主效应的边际模型。综述这些效应。

11.3　参见习题 11.2。进一步的研究发现,在性别和行为类型之间存在交互效应。利用 GEE 法,将相关结构设定为可交换的结构,关于学生具有某一问题行为的概率 π 的模型拟合结果为

$$\text{Logit}(\hat{\pi}) = -0.57 + 1.93S_1 + 0.86S_2 + 0.38R - 0.20G + 0.37G \times S_1 + 0.22G \times S_2,$$

其中 R, G, S_1, S_2 分别表示种族(1 = 白人)、性别(1 = 女性)以及行为类型($S_1 = 1$ 且 $S_2 = 0$ 表示饮酒;$S_1 = 0$ 且 $S_2 = 1$ 表示吸烟;$S_1 = S_2 = 0$ 表示服用大麻)。证明:

a. 所估计的非白人男性吸食大麻的发生比等于 $\exp(-0.57) = 0.57$。

b. 给定性别,所估计的白人具有某种问题行为的发生比是黑人的相应发生比的 1.46 倍。

c. 给定种族,所估计的女性饮酒的发生比是男性的相应发生比的 1.19 倍;所估计的关于吸烟和服用大麻的相应发生比之比分别为 1.02 和 0.82。

d. 给定种族,所估计的女性饮酒(吸烟)的发生比是她吸食大麻的相应发生比的 9.97(2.94)倍。

e. 给定种族,所估计的男性饮酒(吸烟)的发生比是他吸食大麻的相应发生比的 6.89(2.36)倍。对交互项加以解释。

11.4 参见表 11.2。通过对周数进行赋值 (1,2,4), 利用最大似然法或 GEE 法分析该数据。对估计结果加以解释,并将结果与文中使用赋值 (0,1,2) 所得到的结论加以比较。

11.5 以年龄和母亲的吸烟行为作为预测变量,利用边际 Logit 模型分析表 11.9。将结果与第 11.5.5 节中马尔科夫模型的结果进行对比,并加以解释。

11.6 表 11.10 给出的是一个三阶段交叉试验,用来比较安慰剂(干预 A)、小剂量镇痛剂(干预 B)和大剂量镇痛剂(干预 C)对原发性痛经的治疗效果。研究对象被随机分成了六组,分别对应着治疗方式的六种可能次序。在每阶段结束时,每个研究对象对治疗效果进行打分,分为无效(0)或有一定效果(1)两种。令 $y_{i(k)t} = 1$ 表示第 i 个对象在接受第 t 种治疗($t = A,B,C$)后得到缓解,其中第 i 个对象属于第 k 种干预次序($k = 1,\cdots,6$)。假定每种干预次序下的治疗效果是相同的,并令 $\beta_A = 0$, 对模型

$$\mathrm{Logit}[P(Y_{i(k)t} = 1)] = \alpha_k + \beta_t$$

(利用最大似然法或 GEE 法)求 $\{\hat{\beta}_t\}$ 并加以解释。考虑到显著性水平,你会怎样对这些治疗方式进行排序?

表 11.10 习题 11.6 的数据

干预次序	干预结果(A,B,C)							
	000	001	010	011	100	101	110	111
A B C	0	2	2	9	0	0	1	1
A C B	2	0	0	9	1	0	0	4
B A C	0	1	1	8	1	3	0	1
B C A	0	1	1	8	1	0	0	1
C A B	3	0	0	7	0	1	2	1
C B A	1	5	0	4	0	3	1	0

来源:Jones and Kenward(1987)。

11.7 表 11.11 取自堪萨斯州立大学对 262 名养猪农的调查。对于问题"你的家畜疾病信息的主要来源是什么?",回答类别为:(A)专业顾问,(B)兽医,(C)政府或当地推广服务,(D)杂志,(E)饲料公司及其代理人。调查要求被抽中的农民选出所有相关的类别。这个 $2^5 \times 2 \times 4$ 表格给出了按照农民的教育水平(是否至少上过大学)和农场规模(每年出售猪的数量,以千头计)划分的对以上五种信息来源各自的计数(是,否)。

a. 说明为什么下面这种做法是不可取的:对一个将教育水平、农场规模以及信息来源进行交叉划分的 $2 \times 4 \times 5$ 列联表的计数拟合多项分布模型,将信息来源视为结果变量(这个表包括 262 名农夫所给出的 453 个关于信息来源的肯定回答)。

b. 对于教育水平为 i 且农场规模为 s 的农夫,令 $\pi_j(is)$ 表示对第 j 种信息来源回答"是"的概率。表 11.12 给出了通过 GEE 法来拟合不包括教育效应的模型的估计结果,设定操作性相关结构为可交换的结构,

$$\mathrm{Logit}[\pi_j(is)] = \alpha_j + \beta_j s, s = 1,2,3,4。$$

说明如何解释所选择的操作性相关矩阵。说明为什么结果表明,对信息来源 A 具有很强的农场规模的正效应,而对 C,D 和 E 具有大小相似的弱负效应。

c. 限定模型中的参数 $\beta_3 = \beta_4 = \beta_5$, 关于同一斜率的最大似然估计为 -0.184 (SE = 0.063)。说明为什么同时对所有信息来源拟合边际模型要比对每种来源分别拟合模型更好[Agresti 和 Liu(1999)以及 Loughin 和 Scherer(1998)讨论了对这

种形式的数据的分析方法]。

表 11.11 习题 11.7 的数据

			对 D 的回答结果															
			A = 是								A = 否							
			B = 是				B = 否				B = 是				B = 否			
			C = 是		C = 否		C = 是		C = 否		C = 是		C = 否		C = 是		C = 否	
教育水平	售猪数	E	是	否	是	否	是	否	是	否	是	否	是	否	是	否	是	否
未上过大学	<1	是	1	0	0	0	0	0	0	0	2	1	1	2	1	1	5	3
		否	0	0	0	0	0	0	0	1	1	0	0	5	4	7	7	0
	1~2	是	2	0	0	0	0	0	0	0	4	0	0	4	1	0	0	4
		否	0	0	0	0	0	0	0	0	0	0	0	5	0	3	4	0
	2~5	是	3	0	0	0	0	0	0	0	3	0	0	1	2	0	1	1
		否	1	0	0	0	0	0	0	3	0	0	0	2	0	1	4	0
	>5	是	2	0	0	0	0	0	0	0	1	0	1	0	0	1	0	2
		否	1	0	0	2	1	0	1	6	0	1	1	1	0	0	6	0
上过大学	<1	是	3	0	0	0	0	0	0	0	4	0	1	1	0	0	2	11
		否	0	0	0	0	0	0	0	0	4	0	1	2	4	6	14	0
	1~2	是	0	0	0	0	0	0	0	0	2	0	0	1	0	0	1	6
		否	0	0	0	0	1	0	0	1	2	1	0	4	2	7	14	0
	2~5	是	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	3
		否	1	0	0	0	0	0	0	0	0	0	0	5	0	4	4	0
	>5	是	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	2
		否	1	1	0	0	0	1	0	10	0	0	0	4	1	2	4	0

来源：数据由 Kansas State University 的 Tom Loughin 友情提供。

表 11.12 习题 11.7 的输出结果

Working Correlation Matrix

	Col1	Col2	Col3	Col4	Col5
Row1	1.0000	0.0997	0.0997	0.0997	0.0997
Row2	0.0997	1.0000	0.0997	0.0997	0.0997
Row3	0.0997	0.0997	1.0000	0.0997	0.0997
Row4	0.0997	0.0997	0.0997	1.0000	0.0997
Row5	0.0997	0.0997	0.0997	0.0997	1.0000

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Std Error	Z	Pr > |Z|
source	1	-4.4994	0.6457	-6.97	<.0001
source	2	-0.8279	0.2809	-2.95	-0.0032
source	3	-0.1526	0.2744	-0.56	0.5780
source	4	0.4875	0.2698	1.81	0.0708
source	5	-0.0808	0.2738	-0.30	0.7680
size * source	1	1.0812	0.1979	5.46	<.0001
size * source	2	0.0792	0.1105	0.72	0.4738
size * source	3	-0.1894	0.1121	-1.69	0.0912
size * source	4	-0.2206	0.1081	-2.04	0.0412
size * source	5	-0.2387	0.1126	-2.12	0.0341

译者注——source：信息；size * source：规模 * 信息

11.8 参见表 11.13 中对合法流产的态度。对于第 t 个问题($t = 1,2,3$)和性别 g(1 = 女性,0 = 男性)的回答结果 Y_t(1 = 支持,0 = 反对),考虑模型 $\text{Logit}[P(Y_t = 1)] = \alpha + \gamma g + \beta_t$,其中 $\beta_3 = 0$。

表 11.13 习题 11.8 的输出结果

Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	0.8173	0.8173
Row2	0.8173	1.0000	0.8173
Row3	0.8173	0.8173	1.0000

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter	Estimate	Std Error	Z	Pr > |Z|
Intercept	-0.1253	0.0676	-1.85	0.0637
question 1	0.1493	0.0297	5.02	<.0001
question 2	0.0520	0.0270	1.92	0.0544
question 3	0.0000	0.0000	.	.
female	0.0034	0.0878	0.04	0.9688

译者注——Intercept:截距;question:问题;female:女性

a. 使用无结构的相关矩阵进行的 GEE 分析结果如下:它所估计的问题 1 和问题 2 之间的相关系数为 0.826,问题 1 和问题 3 为 0.797,问题 2 和问题 3 为 0.832。这个结果说明选用哪一种操作性相关结构比较合理?

b. 表 11.13 给出了选择可交换的操作性相关结构的 GEE 分析结果,对其中的效应加以解释。

c. 将每个对象的三个回答视为独立的观测值,并进行普通的 logistic 回归,$\hat{\beta}_1 = 0.149$ (SE = 0.066),$\hat{\beta}_2 = 0.052$ (SE = 0.066),$\hat{\gamma} = 0.004$ (SE = 0.054)。用文字说明为什么对象内(within-subject)效应的标准误远远大于 GEE 的相应估计值,而对象间(between-subject)效应的标准误却小于 GEE 估计值。

11.9 参见表 11.7 中的大气污染数据。利用最大似然法或 GEE 法,拟合下列边际 Logit 模型,分别假定:(a)边际同质性,(b)关于时间的线性效应,(c)没有模式。对结果进行解释,并加以比较。

11.10 参见表 12.5 中的临床试验数据,第 12.3.4 节利用随机效应模型对此进行了分析。在此利用 GEE 法来分析这些数据,将每个中心视为一个相关联的群组。

11.11 参考表 10.5。利用有关累积 Logit 的 GEE 法来比较这两个边际分布。将结果与第 10.3.2 节中的最大似然估计结果进行比较。

11.12 参考表 8.19 中关于政府花费的 3^4 表格。利用边际累积 Logit 模型分析这些数据。对模型中的效应加以解释。

11.13 参考表 11.4。

a. 为了在控制结果变量的初始值后对效应进行比较,拟合式 11.7 模型,对入睡时间使用赋值 {10,25,45,75}。另外,拟合包括交互项的模型,并描述拟合的情况(注意:对于前两个基线水平,现行安眠药和安慰剂两种干预方式在跟踪期结果变量具有相似的样本分布;在较高的基线水平,现行安眠药似乎效果更好)。

b. 拟合包括交互项的模型

$$\text{Logit}[P(Y_2 \leqslant j)] = \alpha_j + \beta_1 x + \beta_2 y_1 + \beta_3 x y_1,$$

该模型限定,对使用现行安眠药的一组,效应 $\{\beta_1 x + \beta_2 y_1 + \beta_3 x y_1\}$ 服从模式 $(\tau,\tau,\lambda+\sigma,\lambda)$,使用安慰剂的一组服从 $(\tau,\tau,\sigma,0)$。解释 $\hat{\lambda}$。

11.14　对于表 11.4 的失眠症数据,找出能对该数据进行很好拟合的另一种类型的 Logit 边际模型。解释参数估计的结果,并将所得结果与累积 Logit 模型的结果进行比较。

11.15　参考表 11.9。将数据中母亲吸烟状况的两种水平进行合并。使用一阶马尔科夫链模型能很好地拟合这些数据吗?找出一个对数据拟合充分的对数线性模型。

11.16　利用一个包括前两期结果的转换模型分析表 11.9 的数据。它的拟合情况是不是比第 11.5.5 节中的一阶模型更好?加以解释。

11.17　利用一阶转换模型分析表 11.2 的数据。将所得结果与本章中使用边际模型的结果加以比较。

11.18　表 11.14 所示为一项对学龄儿童的冠状动脉病历险因素的跟踪研究(Woolson and Clarke,1984)。1977 年为 11~13 岁的儿童的一个样本分别在 1977,1979 和 1981 年按照性别和相对体重(肥胖,不肥胖)进行了划分。分析这些数据。

表 11.14　习题 11.18 的数据

性别	结　果[a]							
	NNN	NNO	NON	NOO	ONN	ONO	OON	OOO
男性	119	7	8	3	13	4	11	16
女性	129	8	7	9	6	2	7	14

a NNN 表示在 1977,1979 和 1981 年都不肥胖;NNO 表示在 1977 和 1979 年不肥胖,但在 1981 年为肥胖;依此类推。

来源:授权重印自:the Royal Statistical Society,London(Woolson and Clarke 1984)。

11.19　参考习题 11.7 中有关养猪农的调查数据(表 11.11),建立一个包括所有变量的边际模型来分析这些数据。

11.20　参考习题 7.18 中关于谷类饮食与胆固醇水平的研究(表 7.23)。利用边际模型分析这些数据。

理论与方法

11.21　参考习题 11.1。假设我们将数据表示为关于每个对象的药物与结果之间的 3×2 分表,并利用广义 CMH 方法来进行边际同质性检验。说明为什么 911 + 279 个对每种药物的结果都相同的对象对该检验没有影响。

11.22　令第 i 个对象的第 t 个观测值 $y_{it} = 1$ 或 0, $i = 1,\cdots,n$, $t = 1,\cdots,T$。令 $y_{.t} = \sum_i y_{it}/n$, $y_{i.} = \sum_t y_{it}/T$, 并且 $y_{..} = \sum_i \sum_t y_{it}/nT$。

a. 将 $\{y_{i+}\}$ 视为给定的。假定将 y_{i+} 个"成功"分配给 y_{i+} 个观测值的每种方式都具有同等的可能性。证明 $E(Y_{it}) = y_{i.}$, $\text{var}(Y_{it}) = y_{i.}(1 - y_{i.})$,以及当 $t \neq k$ 时有 $\text{cov}(Y_{it}, Y_{ik}) = -y_{i.}(1 - y_{i.})/(T-1)$(提示:同一行中任意两个单元格之间的协方差都相等,并且由于 y_{i+} 是给定的, $\text{var}(\sum_t Y_{it}) = 0$)。

b. 参考(a)部分。对于大样本情况下的独立对象,说明为什么 $(Y_{.1},\cdots,Y_{.T})$ 近似于多元正态分布,其中两两的相关系数 $\rho = -1/(T-1)$。证明 Cochran 的 Q(Cochran's Q)统计量(Cochran,1950)

$$Q = \frac{n^2(T-1)\sum_{t=1}^{T}(y_{.t} - y_{..})^2}{T\sum_{i=1}^{n} y_{i.}(1 - y_{i.})}$$

近似于 df = $(T-1)$ 的卡方分布(注意:如果 $(X_1,\cdots,X_T)$ 服从具有相同均值和方差的多元正态分布,且每对 (X_t, X_k) 的相关系数都为 ρ, 那么 $\sum(X_t - \bar{X})^2/\sigma^2(1-\rho)$ 服从 df = $(T-1)$ 的卡方分布。在比(a)部分更弱的条件下,Q 的极限也服从卡方分布,有关讨论,参见:Bhapkar 和 Somes(1977))。

c. 证明 Q 不受删除掉 $y_{i1} = \cdots = y_{iT}$ 的观测值的影响。

11.23 考虑模型 $\mu_i = \beta$, $i = 1,\cdots,n$, 并假定 $v(\mu_i) = \mu_i$。如果实际上 $\text{var}(Y_i) = \mu_i^2$, 利用第 11.4 节所介绍的单变量形式的 GEE 法,证明 $u(\beta) = \sum_i(y_i - \beta)/\beta$, 并且 $\hat{\beta} = \bar{y}$。证明式 11.10 中的 V 等于 β/n, 实际的渐近方差(式 11.11)简化为 β^2/n, 对它的一致性估计为 $\sum_i(y_i - \bar{y})^2/n^2$。

11.24 如果实际上 $\text{var}(Y_i) = \mu_i$, 那么假定 $v(\mu_i) = \mu_i^2$,重做习题 11.23。

11.25 对于独立的泊松分布观测值,考虑模型 $\mu_i = \beta$, $i = 1,\cdots,n$。对于 $\hat{\beta} = \bar{y}$, 证明基于模型的渐近方差估计为 $\bar{y}/n$, 而关于渐近方差的稳健估计为 $\sum_i(y_i - \bar{y})^2/n^2$。在下列情况下,你认为哪个更好一些:(a)当泊松模型成立时,(b)当存在严重的过度离散问题时。

11.26 证明式 11.10 等价于由式 4.28 所估计的关于广义线性模型的最大似然估计的大样本协方差公式。

11.27 a. 对于单一的结果变量,类似然(QL)推断与最大似然推断具有怎样的差别?在什么情况下它们会相互等价?

b. 说明从什么意义上来看,GEE 法相当于类似然法的多元形式。

c. 总结类似然法的优缺点。

d. 指出在什么条件下 GEE 参数估计具有一致性,在什么条件下不具有,并对一致性的条件加以说明。

11.28 构建一个与式 11.4 相类似的相邻类别 Logit 或连续比 Logit 模型,对其中的参数加以解释。

11.29 参考第 11.2.3 节最后有关平均入睡时间的分析,说明如何计算分析中所报告的不同均值之差的标准误(注意:其中一个差异使用的是成对样本,而另一个是独立样本)。

11.30 下面这种说法错在哪里:"在一阶马尔科夫链中, Y_t 独立于 Y_{t-2}"。

11.31 假定对数线性模型 $(Y_0, Y_1, \cdots, Y_T)$ 成立,这是一个马尔科夫链吗?

11.32 赌徒 A 和 B 总共有 I 美元。他们在一起重复进行游戏。每次每人都赌 1 美元,胜者得到另一方的钱。每次游戏的结果在统计上独立,A 获胜的概率为 π, B 获胜的概率为 $1-\pi$。当一方获得所有钱时,游戏结束。令 Y_t 表示在 t 次游戏后 A 的总钱数。

a. 证明 $\{Y_t\}$ 服从一阶马尔科夫链。

b. 给出转换概率矩阵(对于这个*赌徒破产*(*gambler's ruin*)问题,0 和 I 是*终结状态*(*absorbing states*)。最终,该链会达到这样一个状态不再变动。所有其他状态都是*暂时的*(*transient*))。

11.33　对于一阶马尔科夫链，如果一步转换概率矩阵都相等，它就具有静态(*stationary or time-homogeneous*)转换概率，即对所有 i 和 j，都满足

$$\pi_{j|i}(1) = \pi_{j|i}(2) = \cdots = \pi_{j|i}(T) = \pi_{j|i}。$$

令 X, Y 和 Z 表示包含 $\{n_{ij}(t), i = 1, \cdots, I, j = 1, \cdots, I, t = 1, \cdots, T\}$ 的 $I \times I \times T$ 表格的三个变量。

a. 如果这个表的期望频数满足对数线性模型 (XY, YZ)，说明为什么所有的转换概率都是静态的(这时，关于静态转换概率的似然比检验统计量等于模型 (XY, YZ) 的拟合优度检验统计量 G^2)。

b. 令 $n_{ij} = \sum_t n_{ij}(t)$。在静态转换概率的假定下，显示如何简化式 11.14 的似然函数，并证明最大似然估计值为

$$\hat{\pi}_{j|i} = n_{ij}/n_{i+}。$$

c. 对于具有静态转换概率的马尔科夫链，令 y_{ijk} 表示在相邻两步中发生的从 i 到 j 再到 k 的转换数量。对于 $\{y_{ijk}\}$，表明对数线性模型 (Y_1Y_2, Y_2Y_3) 的拟合优度是对该链是一阶而不是二阶马尔科夫链的假设检验(Anderson and Goodman, 1957)。

12 随机效应：关于分类结果变量的广义线性混合模型

在第 11 章我们指出，观测值经常存在群组现象。例如，第 i 个群组可能包含关于第 i 个对象的重复测量或者第 i 个家庭中的所有对象。同一群组内的观测值一般会比来自不同群组的观测值更相似，因而，它们之间往往存在正相关。忽略这种组内相关而将组内观测值与组间观测值同等对待的普通分析所给出的标准误是不正确的。

我们在第 11 章集中讨论了关于群组结果变量的*边际*（*marginal*）分布的模型分析，将联合分布的相依结构视为冗余参数。在本章中，我们介绍另外一种方法，该方法将群组层面的项包括进模型里。这些项对属于同一群组的观测值来说取值相同，但在不同群组间的取值不同。如果群组层面的项未被观测到，且将其视为在群组间随机变动时，则称之为*随机效应*（*random effects*）。在第 10.2.4 节中，我们对配对数据进行模型分析时介绍了这种方法。随机效应模型具有*条件*（*conditional*）意义上的解释，当每个群组由一个对象构成时，它被称为*对象别*（*subject-specific*）模型。与之相比，边际模型具有*总体平均*（*population-averaged*）意义上的解释。

关于服从正态分布的结果变量的随机效应模型已经得到了充分发展。相比之下，有关分类变量的随机效应模型只是在最近才开始广泛应用。在本章中，我们将广义线性模型扩展到包括随机效应的情况。第 12.1 节介绍这种扩展的结果——*广义线性混合模型*（*generalized linear mixed model*）。在第 12.2 节中，我们讨论关于二分数据的一个重要特例，*logistic-正态模型*（*logistic-normal model*）。第 12.3 节给出了几个相应的例子。第 12.4 节介绍关于多项分布结果变量的扩展，而第 12.5 节讨论包括多元随机效应的有关模型。在第 12.6 节中，我们探讨随机效应服从正态分布假定的模型的拟合方法。本章中的部分内容取材于：Agresti 等（2000）。

12.1 群组分类数据的随机效应模型

在普通线性回归中，描述变量效应的参数被称为*固定效应*（*fixed effects*）。这同样适用于分类预测变量的所有类别，比如性别、年龄组或治疗方式。与之相对，随机效应通常是针对一个*样本*（*sample*）而言的。例如，对诊所的某一样本进行的研究，模型将属于同一诊所的观测值视为一个群组，并在模型中包括关于每个诊所的随机效应。

广义线性模型对普通回归进行了扩展，它允许非正态分布的结果变量并引入关于均

值的各种连结函数。*广义线性混合模型*(*generalized linear mixed model*,*GLMM*)则进一步扩展了广义线性模型,它允许在线性预测项中同时包括随机效应和固定效应。

12.1.1 广义线性混合模型

令 y_{it} 表示第 i 个群组中的第 t 个观测值,$t=1,\cdots,T_i$。与第 11 章中的 GEE 分析一样,每个群组所包括的观测值数量可以不同。在一项跟踪调查中,即便每个群组具有相同的规模,但是许多群组中可能会出现缺失值。令 $\mathbf{x}_{it}$ 表示由解释变量的取值组成的列向量,其所对应的固定效应模型参数为 $\boldsymbol{\beta}$。令 $\mathbf{u}_i$ 表示关于第 i 个群组的随机效应取值的向量,对于同一群组中的所有观测值来说,随机效应的取值相等。令 $\mathbf{z}_{it}$ 表示由随机效应对应的解释变量组成的列向量。一般来说,随机效应都是单维的。

以 $\mathbf{u}_i$ 为条件,广义线性混合模型与普通的广义线性模型相类似。令 $\mu_{it}=E(Y_{it}\mid\mathbf{u}_i)$。对于连结函数 $g(\cdot)$,广义线性混合模型的线性预测项具有下列形式:

$$g(\mu_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i, \tag{12.1}$$

这里假定随机效应向量服从多元正态分布 $N(\mathbf{0},\boldsymbol{\Sigma})$。其中,协方差矩阵 $\boldsymbol{\Sigma}$ 取决于未知的*方差构成*(*variance components*)以及可能的关联参数。

令 $\mathrm{var}(Y_{it}\mid\mathbf{u}_i)=\phi_{it}v(\mu_{it})$,这里方差函数 $v(\cdot)$ 描述(条件)方差如何取决于均值。与第 4.4 节相同,通常令 $\phi_{it}=1$ 或者 $\phi_{it}=\phi/\omega_{it}$,其中 ω_{it} 是一个已知的权数(如二项分布计数的试验总数),ϕ 是一个未知的离散参数。以 $\mathbf{u}_i$ 为条件,对于所有 i 和 t 来说,模型将 $\{y_{it}\}$ 视为独立的。如同在 10.2.2 节所讨论的那样,就对所有对象求平均的边际分布而言,$\mathbf{u}_i$ 的变动性导致结果变量之间存在非负的关联。这是由于每个群组中的所有观测值都具有共同的随机效应 $\mathbf{u}_i$ 所引起的。

在式 12.1 中,随机效应按照与预测项相同的刻度进入模型。这样表述很方便,同时在许多应用中也很自然。例如,有时随机效应表示由于遗漏了某些解释变量所导致的异质性。考虑单维随机效应以及 $z_{it}=1$ 时的特例。将 u_i 替代为 $u_i^*\sigma$,其中 $\{u_i^*\}\sim N(0,1)$,这时,广义线性混合模型可以表示为:

$$g(\mu_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta} + u_i^*\sigma。$$

它具有普通广义线性模型的形式,其中 $\{u_i^*\}$ 可以看作是某个未观测到的特定协变量的值。因此,随机效应模型与处理未测量的预测变量以及其他形式的缺失数据的方法存在联系。线性预测项中的随机部分反映了那些本应该包括在固定效应部分但却未被包括进来的解释变量所对应的项。另外,有时随机效应也可以代表解释变量中所存在的随机测量误差。如果我们将某个预测变量 x_{it} 替代为 $x_{it}^*+\varepsilon_i$,其中 x_{it}^* 为真值而 ε_i 表示测量误差,那么 ε_i 乘以回归参数所产生的项就被包含在随机效应项中。基于这样的考虑,随机效应也对在普通模型中不包括测量误差所导致的过度离散问题提供了一种解决办法(Breslow and Clayton,1993)。

12.1.2 二分配对数据的 Logit 广义线性混合模型

我们通过一个简单的例子来对广义线性混合模型(式 12.1)进行具体说明,即关于二分配对数据的分析。该数据形成了两个相依的二项分布样本(第 10.1 节)。第 i 个群组由第 i 个配对的回答结果(y_{i1},y_{i2})构成。群组 i 中的第 t 个观测值的取值分别为 $y_{it}=1$(结果为成功)或 0(结果为失败),$t=1,2$。

在第 10.2.2 节中,我们介绍了模型(Cox,1958b;Rasch,1961)

$$\text{Logit}\left[P(Y_{it}=1)\right]=\alpha_i+\beta x_t, \tag{12.2}$$

其中 $x_1=0$ 且 $x_2=1$。对于此模型，β 是一个群组别(cluster-specific)对数发生比之比。在当时，我们将 α_i 视为固定效应，并利用条件最大似然法将其剔除。关于式 12.2 的一个等价的表述形式为：

$$\text{Logit}\left[P(Y_{i1}=1\mid u_i)\right]=\alpha+u_i,\text{Logit}\left[P(Y_{i2}=1\mid u_i)\right]=\alpha+\beta+u_i, \tag{12.3}$$

其中存在某个常数 α，使得 $u_i=\alpha_i-\alpha$。现在，我们将 u_i 视为群组 i 的随机效应，$\{u_i\}$ 相互独立且服从 $N(0,\sigma^2)$ 分布，其中 σ 未知。以 u_i 为条件，我们假定 y_{i1} 和 y_{i2} 相互独立。

式 12.3 模型是式 12.1 模型的一个特例，其中 $\mu_{it}=P(Y_{it}=1\mid u_i)$，$g(\cdot)$ 为 Logit 连结，$\boldsymbol{\beta}'=(\alpha,\beta)$，对所有 i 存在 $\mathbf{x}'_{i1}=(1,0)$ 且 $\mathbf{x}'_{i2}=(1,1)$，以及对所有 i 和 t 存在 $z_{it}=1$。单维随机效应对截距进行了调整，但它没有修正固定效应部分。具有这种形式的随机效应的广义线性混合模型被称为*随机截距*(*random intercept*)模型。与通常所具有的固定截距 α 不同，该模型包括一个随机截距项 $\alpha+u_i$。

令 $Y_1=\sum_i y_{i1}, Y_2=\sum_i y_{i2}$。从边际分布来看，$Y_1$ 服从参数为 $E\{\exp(\alpha+U)/[1+\exp(\alpha+U)]\}$ 的 n 次试验的二项分布，Y_2 服从参数为 $E\{\exp(\alpha+\beta+U)/[1+\exp(\alpha+\beta+U)]\}$ 的二项分布。它们的期望值与 U 有关，其中 U 是服从 $N(0,\sigma^2)$ 分布的随机变量。该模型意味着，在 Y_1 与 Y_2 之间存在一种非负的相关关系，U 的异质性越大(即 σ 越大)，这种关联就越强。当群组所对应的 u_i 取较大的正值时，对于每个 t，$P(Y_{it}=1\mid u_i)$ 的值也相对较大；而当群组所对应的 u_i 取较大的负值时，对于每个 t，$P(Y_{it}=1\mid u_i)$ 的值也相对较小。在这个模型中，只有当 $\sigma=0$ 时，Y_1 和 Y_2 才相互独立。

一个 2×2 的总体平均表格给出了在 $(y_{i1},y_{i2})=(1,1),(1,0),(0,1)$ 或 $(0,0)$ 四种情况下的观测值数量，表中行和列的类别均为(成功，失败)。令 $\{n_{ab}\}$ 表示这些计数。在第 10.1 节曾经分析过的表 12.1 就是这样的一个例子。令 $\{\hat{\mu}_{ab}\}$ 表示式 12.3 模型的边际拟合值。有关模型拟合的讨论，我们将留到第 12.6 节介绍。不过，式 12.3 模型是一种极特殊的情况，即在一个随机效应模型中，固定效应具有封闭形式的最大似然估计，

$$\hat{\beta}=\log(\hat{\mu}_{21}/\hat{\mu}_{12})。$$

当样本的对数发生比之比 $\log(n_{11}n_{22}/n_{12}n_{21})\geq0$ 时，$\{\hat{\mu}_{ab}=n_{ab}\}$，并且 $\hat{\beta}=\log(n_{21}/n_{12})$。这一结果与条件最大似然的估计结果相同(第 10.2.3 节)。Neuhaus 等(1994)证明，只要保证式 12.3 模型的拟合值为 $\{n_{ab}\}$，无论如何选择随机效应分布的参数，上述结论都成立。Lindsay 等(1991)表明，通过一种非参数方法也可以得到这个估计值，详见第 13.2.4 节。该模型隐含着，这个 2×2 表格的真实对数发生比之比至少等于 0。然而，当 $\log(n_{11}n_{22}/n_{12}n_{21})<0$ 时，$\hat{\sigma}=0$，进而拟合值 $\{\hat{\mu}_{ab}=n_{a+}n_{+b}/n\}$ 满足独立性。这时，$\hat{\beta}$ 与式 10.6 边际模型的估计相同，其中 β 等于两个边际分布的 Logit 之差，也即 $\hat{\beta}=\log[(n_{2+}n_{+1})/(n_{1+}n_{+2})]$。

12.1.3 例子：对首相施政表现的评价

对于表 12.1 的数据，在式 12.3 模型中将 $\{u_i\}$ 视为服从正态分布，关于该模型的最大似然拟合为 $\hat{\beta}=\log(86/150)=-0.556(\text{SE}=0.135)$，其中 $\hat{\sigma}=5.16$。这一结果与条件最大似然估计(式 10.10)相同，其标准误为 $[(1/86)+(1/150)]^{1/2}$。因而，对于一个给定的对象，所估计的在第二次调查中持支持态度的发生比是第一次的相应发生比的

exp(−0. 556) = 0. 57 倍。结果中 $\hat{\sigma}$ 的值较大表明,两次调查结果之间具有很强的关联,其样本发生比之比高达 35. 1。

表 12. 1 对首相施政表现的评价

第 1 次调查	第 2 次调查		小 计
	支持	不支持	
支持	794	150	944
不支持	86	570	656
小计	880	720	1 600

12.1.4 扩展:Rasch 模型与项目反应理论

对 Logit 配对模型(式 12. 3)进行扩展,允许在每个群组中包括 $T>2$ 个观测值。这时,随机截距模型可以表示为:

$$\text{Logit}\left[P(Y_{it}=1 \mid u_i)\right] = u_i + \beta_t, \tag{12.4}$$

其中 $\{u_i\}$ 相互独立且服从 $N(0,\sigma^2)$ 分布。等价地,模型中可以加入一个截距项 α 或令 $E(u_i)=\alpha$,但是这时模型的可识别性要求参数满足一定的约束条件如 $\beta_T=0$。

这个广义线性混合模型早期主要应用于心理测量学方面。该模型描述了在测验中对一组共 T 个问题的回答情况。第 i 个对象对第 t 个问题回答正确的概率 $P(Y_{it}=1 \mid u_i)$ 取决于他本人的综合能力,由 u_i 来表示,以及该问题的难易程度,由 β_t 来表示。这样的模型被称为*项目反应模型*(*item-response models*)。项目反应模型的 Logit 形式(式 12. 4)被称为 *Rasch 模型*(*Rasch model*)(Rasch, 1961)。在估计 $\{\beta_t\}$ 时,Rasch 将 $\{u_i\}$ 视为固定效应并使用了条件最大似然法,对此方法我们在第 10. 2. 3 节有关配对数据的分析中已经介绍过。后来的研究者对该模型使用正态随机效应方法,并在模型中选用 probit 连结(如:Bock and Aitkin, 1981)。

Rasch 模型中的 $\{\beta_t\}$ 与相应边际模型如式 11. 1 中的参数不同,因为这里的效应是对象别效应。Rasch 模型对应的是一个由回答结果与研究对象所形成的 $T\times 2\times n$ 表格,而边际模型对应的则是一个由回答结果的 T 个边际分布所形成的 $T\times 2$ 表格,它将所有对象进行了合并。在式 12. 4 模型中,对于给定的对象 i,它的第 s 和第 t 个观测值之间存在

$$\beta_s - \beta_t = \text{Logit}\left[P(Y_{is}=1 \mid u_i)\right] - \text{Logit}\left[P(Y_{it}=1 \mid u_i)\right],$$

这是一个以该对象为条件的对数发生比之比。相反,在式 11. 1 边际模型中,相应的总体平均效应为

$$\beta_s - \beta_t = \text{Logit}\left[P(Y_{hs}=1)\right] - \text{Logit}\left[P(Y_{it}=1)\right],$$

它反映的是随机选取的对象 h 的第 s 个观测值与随机选取的对象 i 的第 t 个观测值(即,h 和 i 是*独立的*(*independent*)观测值)之间的对数发生比之比。

12.1.5 随机效应模型与条件最大似然法的比较

假定我们将式 12. 4 模型中的 $\{u_i\}$ 视为固定效应,而不是随机效应,那么,考虑关于 $\{\beta_t\}$ 和 $\{u_i\}$ 的普通最大似然估计。随着 n 的增加,参数的数量也会增加,因为每个对象都对应着一个 u_i。这样,即使 $\{\beta_t\}$ 的数量不随着 n 的增加而增加,它的最大似然估计值 $\{\hat{\beta}_t\}$ 仍然不具有一致性。当参数数量与研究对象的数量具有相似的数量级时,在许多模

型中都会发生这种情况。最大似然估计值的渐近最优特性,比如一致性,要求在 n 增加时参数的数量是固定的。对于式 12.4 模型,关于 $\{\beta_t\}$ 的最大似然估计的偏误数量级为 $T/(T-1)$(Anderson,1980,pp. 244-245)。例如,在式 12.2 配对模型中,依概率 $\hat{\beta}\rightarrow 2\beta$(习题 10.24)。

由于这一原因,拟合固定效应模型的较好方法是*条件*(*conditional*)最大似然法。以 $\{u_i\}$ 的充分统计量 $\{S_i=\sum_t y_{it}, i=1,\cdots,n\}$ 为条件,可以将其消除。在项目反应模型中,这些充分统计量就是每个对象回答对的问题数量。以 $\{S_i\}$ 为条件,$\{y_{it}\}$ 的分布独立于 $\{u_i\}$。这时,最大化相应的似然函数可以得到关于 $\{\beta_t\}$ 的一致估计。这是对第 10.2.3 节中关于配对数据的对象别 logistic 模型(式 10.8)的扩展。有关细节,参见:Anderson(1980)。

与随机效应方法相比,条件最大似然法具有一定的优点。我们不需要假定关于 $\{u_i\}$ 的参数分布,而在随机效应模型中,往往很难去验证相应分布的假定是否正确。另外,条件最大似然法还适用于回顾性样本的研究。在这种情况下,随机效应方法可能会产生偏误,因为群组不再是一个随机样本(Neuhaus and Jewell, 1990b)。

然而,条件最大似然法也存在严重的不足。它只局限于典型连结(Logit),只有这种情况下关于 $\{u_i\}$ 的充分统计量才存在。更重要的是,如第 10.2.7 节所述,它只局限于对组内固定效应的推断。条件法消除了模型中估计组间效应所需要的变动性,相应的情况我们将在下文考虑。另外,条件最大似然法不提供关于 $\{u_i\}$ 的信息,比如对它们取值的预测、变动性的估计或者它们所决定的概率。最后,在包括协变量的更具一般性的模型中,与随机效应方法相比,条件最大似然法在估计固定效应时会损失一些效率(参见注解 12.2)。

12.2 二分结果变量:Logistic-正态模型

包含随机截距的项目反应模型(式 12.4)是一类重要的二分数据随机效应模型的特例,该类模型被称为 *logistic-正态模型*(*logistic-normal models*)。在包括单维的随机效应的情况下,该模型可以表示为:

$$\text{Logit}\left[P(Y_{it}=1\mid u_i)\right]=\mathbf{x}'_{it}\boldsymbol{\beta}+u_i, \tag{12.5}$$

其中 $\{u_i\}$ 是服从 $N(0,\sigma^2)$ 的独立变量。它是广义线性混合模型(式 12.1)的一个特例,这里 $g(\cdot)$ 为 Logit 连结,且随机效应的结构简化为一个随机截距。Logistic-正态模型具有很长的历史,至少可以追溯到 Cox(1970,书中的习题 20)关于配对数据的分析模型(式 12.3),另见:Pierce 和 Sands(1975)。

在更为一般的情况下,式 12.5 模型中的连结函数可以是任意一个累积分布函数的反函数。在这样的模型中,Y_{is} 和 Y_{it} 被视为(给定 u_i)条件独立,但是它们的边际分布却存在非负相关。令 Φ 表示连结函数的反函数,即累积分布函数,那么,对于任意的 $s\neq t$,

$$\begin{aligned}\text{cov}(Y_{is},Y_{it}) &= E\left[\text{cov}(Y_{is},Y_{it}\mid u_i)\right]+\text{cov}\left[E(Y_{is}\mid u_i),E(Y_{it}\mid u_i)\right]\\ &= 0+\text{cov}\left[\Phi(\mathbf{x}'_{is}\boldsymbol{\beta}+u_i),\Phi(\mathbf{x}'_{it}\boldsymbol{\beta}+u_i)\right]。\end{aligned} \tag{12.6}$$

上式最后一项协方差中的函数都随着 u_i 单调递增,因而它们存在非负相关。当预测变量 $\mathbf{x}$ 的取值在每个 t 处相同时,这个模型的联合分布是可交换的。对于群组数据来说,这是很现实的情况。但是,在跟踪研究中,时间上更接近的观测值往往具有更强的相关关系。

通常来说,应用广义线性混合模型的主要目的是对固定效应进行统计推断,模型的

随机效应部分用来表示同一群组内的观测值之间的正相关是如何发生的。然而,与随机效应有关的参数本身有时也是研究关注的重点。例如,关于随机截距标准差的估计 $\hat{\sigma}$ 可能是反映总体异质性程度的一个重要指标。

12.2.1 解释 Logistic-正态模型中的异质性

当 $\sigma=0$ 时,logistic-正态模型(式 12.5)就简化为将所有观测值都视为独立观测值的普通 logistic 回归模型。而当 $\sigma>0$ 时,我们应当怎么解释这个模型所隐含的效应的变动性呢?

考虑在预测项取值为 $\mathbf{x}_{it}$ 的观测值 y_{it} 与预测项取值为 $\mathbf{x}_{hs}$ 的观测值 y_{hs}。它们的对数发生比之比等于

$$\text{Logit}\left[P(Y_{it}=1\mid u_i)\right]-\text{Logit}\left[P(Y_{hs}=1\mid u_h)\right]=(\mathbf{x}_{it}-\mathbf{x}_{hs})'\boldsymbol{\beta}+(u_i-u_h)。$$

我们无法观测到 (u_i-u_h),它服从 $N(0,2\sigma^2)$ 分布。不过,该对数发生比之比有 $100(1-\alpha)\%$ 的概率落入区间

$$(\mathbf{x}_{it}-\mathbf{x}_{hs})'\boldsymbol{\beta}\pm z_{\alpha/2}\sqrt{2}\sigma。\qquad(12.7)$$

当 $\sigma=0$ 时,$(\mathbf{x}_{it}-\mathbf{x}_{hs})'\boldsymbol{\beta}$ 是不包括随机效应的普通模型所对应的对数发生比之比。当 $\sigma>0$ 时,$(\mathbf{x}_{it}-\mathbf{x}_{hs})'\boldsymbol{\beta}$ 是两个属于同一群组($h=i$)或具有相同的随机效应值的观测值之间的对数发生比之比。假定对于来自不同群组的观测值有 $\mathbf{x}_{it}=\mathbf{x}_{hs}$,那么,利用 $z_{0.25}=0.674$,对数发生比之比取值分布的中间 50% 落入区间 $\pm0.674\sqrt{2}\sigma=\pm0.95\sigma$。因此,具有较大的随机效应的观测值与具有较小随机效应的观测值之间的中位数发生比之比(median odds ratio)等于 $\exp(0.95\sigma)$。当只有一个预测变量且 $x_{it}-x_{hs}=1$ 时,中位数发生比之比等于 $\exp(\beta+0.95\sigma)$。Larsen 等(2000)对此给出了相应的解释。

12.2.2 条件模型与边际模型的联系

在广义线性混合模型中,固定效应参数 $\boldsymbol{\beta}$ 实际上是给定随机效应后的条件参数。固定效应可以分为两种情况。第一种,考虑一个解释变量对同一群组中的观测值可以取不同值的情况。例如,在一项比较 T 种药品的交叉研究中,在对某个对象进行 T 次观测所形成的群组中,每次使用的药物都不同。对于这种解释变量,模型中它所对应的参数指的是,在群组内(即对象别)该预测变量每增加一个单位对结果变量的影响。这时候,模型中的随机效应和其他解释变量被认为是给定的。这种解释变量的效应是一种"组内"或"对象内"效应。

第二种,考虑一个解释变量对同一群组内的观测值取值相同的情况。当以每个对象作为群组时,对象的性别就是这样的一个例子。对于这种解释变量,它的回归系数指的是,在群组间该预测变量每增加 1 单位对结果变量的影响。例如,通过虚拟变量来比较不同性别的对象。但是,在广义线性混合模型中,只有当两个组的随机效应(以及模型中的其他解释变量)取相同的值时,这种固定效应才适用:比如,比较具有相同随机效应的男性和女性。

正是在这个意义上说,随机效应模型是条件模型,组内和组间效应都以随机效应的取值为条件。相反,边际模型中的效应是对所有群组的平均(即,总体平均),因此这些效应并不是在随机效应的某个固定取值处的比较。事实上,这两类模型的一个根本区别在于,当连结函数是非线性函数时,比如 Logit 连结,边际模型中的总体平均效应往往小于

广义线性混合模型中的群组别(cluster-specific)效应。

具体地说,广义线性混合模型(式 12.1)对应的是条件均值,$\mu_{it}=E(Y_{it}|\mathbf{u}_i)$。通过连结函数的反函数,可得:

$$E(Y_{it}\mid\mathbf{u}_i)=g^{-1}(\mathbf{x}'_{it}\boldsymbol{\beta}+\mathbf{z}'_{it}\mathbf{u}_i)。$$

从边际分布的角度来看,对随机效应求平均,有

$$E(Y_{it})=E\left[E(Y_{it}\mid\mathbf{u}_i)\right]=\int g^{-1}(\mathbf{x}'_{it}\boldsymbol{\beta}+\mathbf{z}'_{it}\mathbf{u}_i)f(\mathbf{u}_i;\boldsymbol{\Sigma})d\mathbf{u}_i,$$

其中$f(\mathbf{u};\boldsymbol{\Sigma})$是关于随机效应的$N(\mathbf{0},\boldsymbol{\Sigma})$密度函数。在连结函数为恒等连结的情况下,

$$E(Y_{it})=\int(\mathbf{x}'_{it}\boldsymbol{\beta}+\mathbf{z}'_{it}\mathbf{u}_i)f(\mathbf{u}_i;\boldsymbol{\Sigma})d\mathbf{u}_i=\mathbf{x}'_{it}\boldsymbol{\beta}。$$

这时,相应边际模型的模型形式和效应 $\boldsymbol{\beta}$ 与之相同。对于其他连结函数,这一点并不成立。例如,在 logistic-正态模型(式 12.5)中,

$$E(Y_{it})=E\left[\frac{\exp(\mathbf{x}'_{it}\boldsymbol{\beta}+u_i)}{1+\exp(\mathbf{x}'_{it}\boldsymbol{\beta}+u_i)}\right]。$$

除非 u_i 服从退化分布(degenerate distribution,$\sigma=0$),该期望不具有 $\exp(\mathbf{x}'_{it}\boldsymbol{\beta})/[1+\exp(\mathbf{x}'_{it}\boldsymbol{\beta})]$的形式。

不过,这两类模型的估计结果之间存在着一种近似关系。当 logistic-正态模型中的效应为 $\boldsymbol{\beta}$ 并且 σ 较小时,Zeger 等(1988)表明,

$$E(Y_{it})\approx\exp(c\mathbf{x}'_{it}\boldsymbol{\beta})/\left[1+\exp(c\mathbf{x}'_{it}\boldsymbol{\beta})\right],\tag{12.8}$$

其中 $c=[1+0.6\sigma^2]^{-1/2}$。由于边际模型中的效应约为条件模型中相应效应的 c 倍,因此它的绝对值一般较小,二者之间的差异随着 σ 的增加而增加。当 $\boldsymbol{\beta}$ 接近于 0 时,Neuhaus 等(1991)表明,边际模型的效应近似等于 $\boldsymbol{\beta}(1-\rho)$,其中在 $\boldsymbol{\beta}=0$ 时 ρ 等于 $\mathrm{corr}(Y_{it},Y_{is})$。同样,二者之间的差异随着 σ 的增加而增加,因为 ρ 是 σ 的增函数。

对于表 12.1 中有关首相施政满意度的数据,式 12.3 模型的最大似然估计为 $\hat{\beta}=-0.556$,关于$\{\mu_i\}$的变动性的估计为 $\hat{\sigma}=5.16$。根据式 12.8 的近似,相应的边际模型的参数估计约等于$[1+0.6(5.16)^2]^{-1/2}(-0.556)=-0.135$。对于该数据,实际的边际效应估计值等于样本边际分布的对数发生比之比,即

$$\log\left[(880/720)/(944/656)\right]=-0.163。$$

事实上,边际效应比条件效应小得多,但是当 $\hat{\sigma}$ 较小时,两种估计值之间的这种近似关系会更密切。当 $\beta=0$ 时,相当于对以上数据拟合一个对称性模型,其中 $\hat{\mu}_{12}=\hat{\mu}_{21}=(n_{12}+n_{21})/2$。该 2×2 表格的相关系数等于 0.699,据此,条件估计 -0.556 对应的边际估计应为 $-0.556(1-0.699)=-0.167$,与实际值 -0.163 非常接近。

图 12.1 展示了为什么边际效应会小于条件效应。在只有一个解释变量 x 的情况下,该图给出了几个存在明显异质性的对象对应的对象别曲线 $P(Y_{it}=1|u_i)$,即随机效应的 σ 相对较大的情况。在 x 取任意给定值时,条件均值 $E(Y_{it}|u_i)=P(Y_{it}=1|u_i)$都存在变动性。对这些条件均值求平均就得到了边际均值 $E(Y_{it})$。在 x 的不同取值处分别求平均的结果就是图中的这条虚线。它的斜率更小,事实上,它的形状并不与 logistic 曲线完全一致。在其他的广义线性混合模型中,也存在类似的现象。然而,对于使用 probit 连结的二分数据,具有正态随机效应的条件 probit 模型确实对应着一个 probit 形式的边际模型(习题 12.29)。在单维随机截距模型中,probit 连结的边际效应等于相应的条件效应

乘以$[1+\sigma^2]^{-1/2}$(Zeger et al., 1988)。在对数线性形式的广义线性混合模型中,有关条件模型与边际模型之间的联系,我们将在第 13. 5. 1 节探讨。

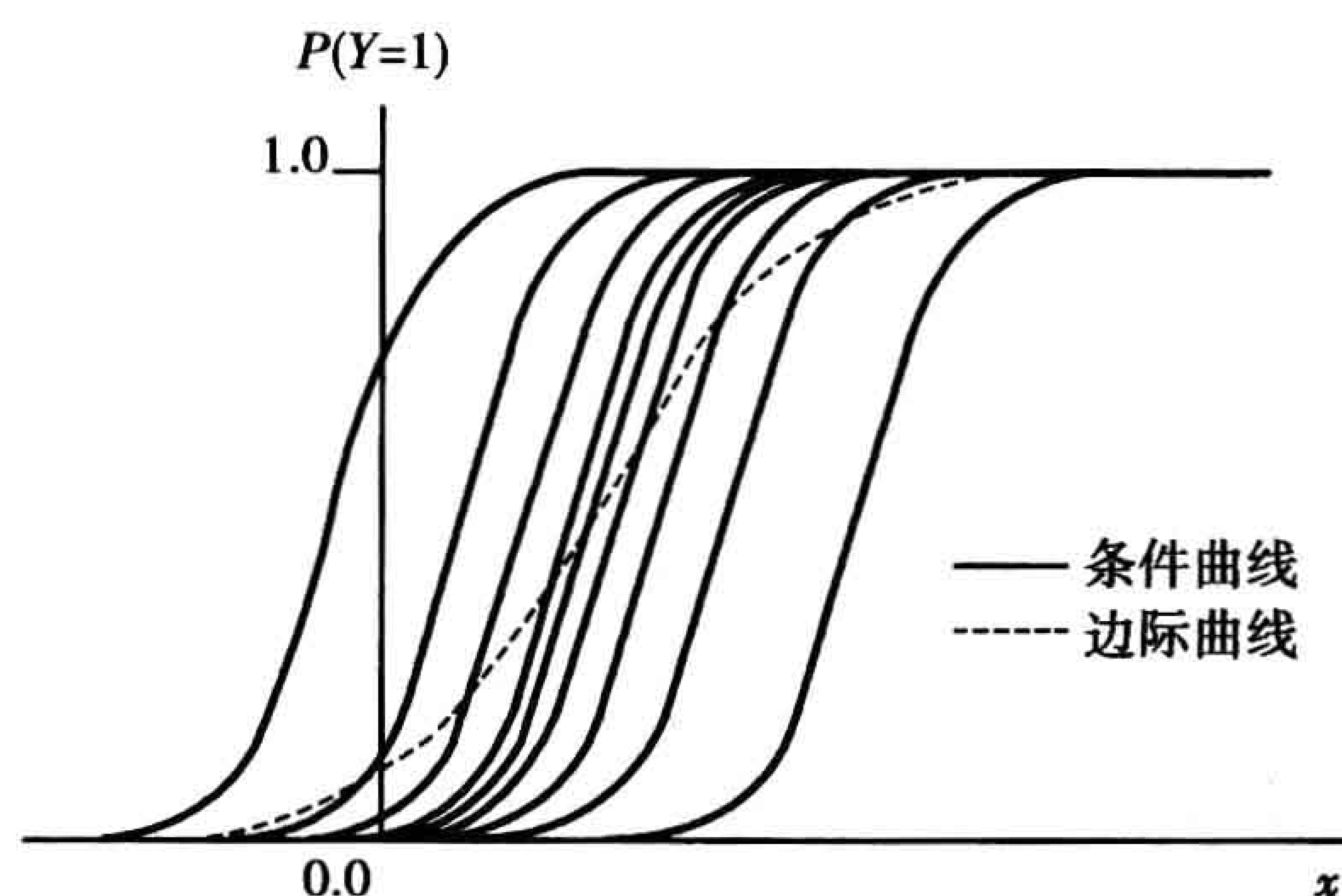


图 12. 1 Logistic 随机截距模型,图中显示的是条件(对象别)曲线以及求平均后的边际(总体平均)曲线

12. 2. 3 有关条件模型和边际模型的评论

随机效应模型描述的是条件(对象别)效应,而边际模型描述的是总体平均效应。不同统计学家偏好不同的模型,但总体上,大多数人都觉得两种模型都有价值,使用哪一种取决于实际应用中的具体情况。

如果研究者想设定组内观测值之间正相关的形成机理、估计组别效应、估计群组效应的变动性或者对联合分布进行分析,那么就应该选择条件模型方法。用来决定模型形式的潜变量理论(比如,第 6. 6. 1 节中关于二分模型的容差(tolerance)理论,习题 6. 28 中的临界(threshold)理论,以及习题 6. 29 中的效用(utility)理论),应用在群组层面而不是边际层面更加自然。给定一个条件模型,我们可以从中推出有关边际分布的信息。也就是说,条件模型隐含着边际模型,但是边际模型本身并不能决定条件模型(尽管注解 12. 10提到了二者之间的内在联系)。

在许多调查或流行病研究中,研究目标之一是比较在不同人群之间某种结果发生的相对频繁程度。这时,研究主要关注的数字是不同群体的边际概率之间的组间发生比之比,也即,所关注的是组间效应,而不是组内效应。当边际效应是研究的主要关注点时,一般直接对边际分布进行模型分析更加简便,也可能更适当。在这种情况下,我们可以对模型进行参数设定,以使得回归参数具有直接的边际解释。而对决定这些边际分布的联合分布构建一个更具体的模型,如随机效应模型,增大了出现模型设定错误的可能性。例如,在跟踪数据中,给定随机效应后观测值之间相互独立的假定不一定必然成立。我们在第 11 章指出,在拟合边际模型时,尽管某些情况下最大似然估计是可行的,但 GEE 法在运算上更为简单而且适用范围也更广。GEE 法的一个缺点在于,它不将随机效应明确地纳入模型,因而也不估计这些效应。另外,由于没有设定关于结果变量的联合分布,基于似然的统计推断对 GEE 法来说不适用。

在第 12. 2. 2 节我们指出,条件效应通常大于边际效应,并且二者的差别会随着方差构成项取值的增加而上升。不过,一般来说,效应的显著性水平(如估计值与其标准误的比率)在两种模型中还是相似的。如果在条件模型中某个效应比另一个更重要,在边际模型中通常也会出现相同的情况。因此,模型的选择基本不会影响统计推断的结论。

然而,需要特别指出的是,由于边际模型中效应的大小取决于条件模型中的异质性

程度。在比较方差构成差别很大的两组或两个变量的效应时,边际模型与条件模型中效应的相对大小可能并不一致。由式 12.8 可知,在二分数据情况下,对具有较大的方差构成的组来说,条件效应与边际效应之间的差别也更大。例如,假定存在两个组别,青年组和老年组,两组在一项比较两种药物疗效的交叉研究中具有相同的条件效应。如果结果变量在老年组的分布存在更大的异质性,那么该组的边际效应可能会比青年组小。尽管二者的条件效应相等,只是因为老年组具有更大的方差构成,它们的边际效应却不同。在这种情况下,条件效应可能更有意义。

最后,无论边际模型还是条件模型,缺失数据是多元结果变量分析中的一个常见问题。除非数据缺失是随机的,否则,最大似然推断的结果可能会出现偏误。GEE 法通常要求数据缺失完全随机(第 11.4.5 节)的更强条件。因此,对缺失机理本身的分析或通过灵敏度分析来检验缺失数据的可能影响,是数据分析的一个重要组成部分。

无论选择哪种模型形式,对统计学家来说,即便向应用者解释清楚为什么在非线性连结函数的情况下边际效应与条件效应不相等都是一个挑战。诸如图 12.1 的示例在这方面会有所帮助。Neuhaus(1992)以及 Pendergast 等(1996)系统回顾了有关群组二分数据的分析方法,包括条件模型和边际模型。Agresti 和 Natarajan(2001)综述了有关群组定序数据的条件模型和边际模型分析。

12.3 二分数据随机效应模型的例子

在接下来的三节中,我们介绍几个关于随机效应模型的例子。本节首先考虑二分结果变量的情况。

12.3.1 二项分布比例的小区域估计

小区域估计(*small-area estimation*)指的是对数量众多的地理区域进行参数估计,而每个区域只包括相对很少的观测值。例如,有人或许想知道如失业率或拥有健康保险的家庭比例等方面的分县估计数据。对于一项全国或全州的调查来说,在一个县中可能只包括几个观测值。这时,用样本比例来估计该县的真实比例可能非常糟糕。将每个县作为群组的随机效应模型可以给出更精确的估计。该拟合过程通过假定真实的比例按照某种分布变动,"借用了整个样本的信息",也即在估计某个给定县的比例时,模型利用了所有县的数据。

令 π_i 表示第 i 个区域的真实比例,$i = 1,\cdots,n$。这些区域可以是所关注的总体,也可以是总体的一个样本。令 $\{y_i\}$ 表示独立的 $\text{bin}(T_i,\pi_i)$ 变量,也即,$y_i = \sum_{t=1}^{T_i} y_{it}$,其中 $\{y_{it}, t = 1,\cdots,T_i\}$ 相互独立,并有 $P(Y_{it} = 1) = \pi_i$ 以及 $P(Y_{it} = 0) = 1 - \pi_i$。样本比例 $\{p_i = y_i/T_i\}$ 是固定效应模型

$$\text{Logit}(\pi_i) = \alpha + \beta_i \quad (i = 1,\cdots,n)$$

对 $\{\pi_i\}$ 的最大似然估计。这个模型是饱和模型,用 n 个非冗余参数(模型的约束条件可以设如 $\sum_i \beta_i = 0$)描述 n 个二项分布观测值。

在 $\{T_i\}$ 很小的情况下,$\{p_i\}$ 的标准误非常大。因此,$\{p_i\}$ 表现出的变动性可能比 $\{\pi_i\}$ 大得多,尤其是当 $\{\pi_i\}$ 取值相似时。这时,一种有意义的做法是使 $\{p_i\}$ 向其总体均值收缩。这一点可以通过随机效应模型

$$\text{Logit}\left[P(Y_{it} = 1 \mid u_i)\right] = \alpha + u_i \tag{12.9}$$

来实现,其中$\{u_i\}$是独立的$N(0,\sigma^2)$变量。这个模型是单维随机效应方差分析所对应的Logit形式。当$\sigma=0$时,所有π_i都相等。

在这个模型中,

$$\hat{\pi}_i = \exp(\hat{\alpha}+\hat{u}_i)/\left[1+\exp(\hat{\alpha}+\hat{u}_i)\right]。$$

模型的估计值与样本比例p_i不同。如果$\hat{\sigma}=0$,那么所有的$\hat{u}_i=0$。这时,关于每个π_i的随机效应估计等于$(\sum_{i=1}^{n}\sum_{t=1}^{T_i}y_{it})/(\sum_i T_i)$,即将所有$n$个样本合并后的总的样本比例。当所有$\pi_i$确实都相等时,这个对共同值的估计比单个样本的样本比例好得多。

一般来说,随机效应模型估计值将单个样本比例向总的样本比例收缩。收缩的程度随着$\hat{\sigma}$的增加而下降;收缩幅度也会随着$\{T_i\}$的增大而下降,随着每个样本包括的数据增多,我们对单个的样本比例的信任度也会上升。所预测的随机效应$\hat{u}_i$是给定数据后所估计的u_i分布的均值(参见第12.6.7节)。这个预测取决于所有数据,而不只是第i个区域的数据。这样做的一个好处是,可以减小对真实值进行估计时所产生的均方差(mean-squared error)。

我们通过对1996年美国总统大选进行民调的一个模拟样本来展示式12.9模型,样本规模为2 000。对于第i个州($i=1,\cdots,51$,其中$i=51$表示哥伦比亚特区)的T_i个观测值,y_i服从$\text{bin}(T_i,\pi_i)$,这里π_i是在1996年大选中以投票给克林顿或者共和党候选人鲍勃·多尔为条件,第i个州投票给比尔·克林顿的实际比例。在本例中,T_i与各州的人口总数成比例,并且$\sum_i T_i=2\ 000$。表12.2给出了$\{T_i\}$、$\{\pi_i\}$以及$\{p_i=y_i/T_i\}$。

对式12.9模型的最大似然拟合结果为$\hat{\alpha}=0.163$,$\hat{\sigma}=0.29$。可以通过随机效应的预测值(由SAS的PROC NLMIXED给出),计算相应的比例估计值$\{\hat{\pi}_i\}$,如表12.2所示。由于大部分$\{T_i\}$都很小,而且$\hat{\sigma}$也相对较小,这些估计值反映了从样本比例向支持克林顿的总比例(0.548)的明显收缩。$\{\hat{\pi}_i\}$的变动范围为0.468(TX,德克萨斯)到0.696(NY,纽约),而样本比例的变动范围是从0.111(爱达荷)到1.0(DC,哥伦比亚特区)。其中,基于较少观测值的样本比例收缩的幅度更大,如哥伦比亚特区。尽管包括了随机效应的模型估计值相对而言更具同质性,它们却比这些样本比例更接近于真实值。

表12.2 在1996年美国总统大选中以投票给克林顿或者多尔为条件,所估计的投票给克林顿的比例[a]

州	T_i	π_i	p_i	$\hat{\pi}_i$	州	T_i	π_i	p_i	$\hat{\pi}_i$
AK	5	0.394	0.200	0.508	FL	108	0.532	0.602	0.583
AL	32	0.463	0.500	0.524	GA	56	0.494	0.554	0.548
AR	19	0.594	0.526	0.537	HI	9	0.643	0.556	0.543
AZ	34	0.512	0.618	0.573	IA	22	0.557	0.500	0.528
CA	240	0.572	0.538	0.538	ID	9	0.391	0.111	0.472
CO	29	0.492	0.586	0.558	IL	89	0.596	0.539	0.540
CT	25	0.604	0.720	0.602	IN	44	0.468	0.432	0.488
DC	4	0.903	1.000	0.576	KS	19	0.400	0.316	0.477
DE	5	0.586	0.400	0.527	KY	29	0.506	0.448	0.506

续表

州	T_i	π_i	p_i	$\hat{\pi}_i$	州	T_i	π_i	p_i	$\hat{\pi}_i$
LA	33	0.566	0.667	0.592	OH	84	0.536	0.488	0.507
MA	46	0.686	0.739	0.637	OK	23	0.456	0.478	0.520
MD	38	0.586	0.474	0.511	OR	24	0.547	0.625	0.569
ME	9	0.627	0.778	0.578	PA	90	0.552	0.567	0.558
MI	73	0.573	0.589	0.570	RI	7	0.689	0.571	0.545
MN	35	0.594	0.571	0.554	SC	28	0.469	0.571	0.552
MO	41	0.535	0.561	0.550	SD	6	0.479	0.667	0.555
MS	21	0.472	0.333	0.477	TN	40	0.513	0.500	0.522
MT	7	0.483	0.429	0.526	TX	144	0.473	0.444	0.468
NC	55	0.475	0.455	0.494	UT	15	0.380	0.333	0.490
ND	5	0.461	0.600	0.546	VA	51	0.489	0.412	0.473
NE	13	0.395	0.462	0.524	VT	4	0.633	0.500	0.538
NH	9	0.567	0.556	0.543	WA	42	0.572	0.619	0.578
NJ	60	0.600	0.667	0.611	WI	39	0.559	0.487	0.517
NM	13	0.540	0.462	0.524	WV	14	0.584	0.571	0.548
NV	12	0.506	0.500	0.533	WY	4	0.426	0.250	0.518
NY	137	0.660	0.752	0.696					

a π_i，真实值；p_i，样本比例；$\hat{\pi}_i$，随机效应模型的估计值。

12.3.2 对重复测量的二分结果变量的模型分析

在第 12.1.4 节中，我们介绍了具有随机效应形式的 Rasch 模型，其结果变量为重复测量的二分变量。这个模型可以扩展到包括协变量的情况。

我们以表 10.13 的数据为例来加以展示，这些数据曾在第 10.7.2 节分析过。调查对象分别回答了在三种不同情形下是否支持合法流产的问题。另外，表 10.13 还按照性别对调查对象进行了分组。令 y_{it} 表示第 i 个对象对第 t 个题目的回答，其中 $y_{it}=1$ 代表支持。考虑模型

$$\text{Logit}\left[P(Y_{it}=1 \mid u_i)\right] = u_i + \beta_t + \gamma x_i, \tag{12.10}$$

其中 $x_i=1$ 代表女性、0 代表男性，$\{u_i\}$ 服从独立的 $N(0,\sigma^2)$（等价地，我们可以对 $\{\beta_t\}$ 加以限定，而在模型中包括截距 α）。这里，假定性别效应 γ 对每个题目都相同，并且 $\{\beta_t\}$ 是关于这些题目的参数。

由于式 12.10 模型意味着对各个题目的回答之间存在非负关联，这种情况下我们应当使用项目和量表。在研究有关合法流产的态度时，给定回答类别为（是，否），如果一个问题问“您赞同法律应当允许未婚妇女进行流产吗?”，而另一个问题却问“您赞同法律应当禁止在怀孕的最后三个月进行流产吗?”，这种问法是不恰当的。

表 12.3 给出了相应的最大似然拟合结果。$\{\hat{\beta}_t\}$ 之间的差异表明，对题目 1（当家庭收入很低而无法抚养更多的孩子时，法律应当允许进行流产）的支持度高于另两个题目，有微弱的证据显示对题目 2（当妇女未婚并且不想嫁给胎儿的父亲时）的支持度稍高于题目 3（无论什么原因）。通过对数发生比之比可以解释固定效应的估计结果。例如，对于给定的对象，不论性别，所估计的支持题目 1 的发生比是支持题目 3 的相应发生比的

exp(0.83) = 2.3 倍。由于 $\hat{\gamma} = 0.01$,在具有相似随机效应取值的男性和女性之间,所估计的支持合法流产的概率对每个题目都不存在明显差异。

在这些数据中,调查对象之间存在着高度异质性($\hat{\sigma} = 8.6$)。因此,对三个题目的回答之间具有很强的关联。这一点也表现为在 1 850 个对象中有多达 1 595 人对三个题目的回答完全一致:即,回答结果为(0,0,0)或者(1,1,1)。这意味着,对象之间的发生比之比具有极大的变动性。由式 12.7 可知,对于给定性别的不同对象,所估计的题目 1 和题目 3 的发生比之比的中间 50% 取值在 exp(0.83 − 0.95 × 8.6) 和 exp(0.83 + 0.95 × 8.6) 之间变动。

对于列联表数据,我们可以计算模型对单元格的拟合值。这需要对所估计的随机效应的分布求积分,以得出每种结果组合的边际概率。在关于参数的最大似然估计中,由于在给定 u_i 后各个结果变量之间是相互独立的,结果组合($y_{i1}, \cdots, y_{iT}$) 发生的概率等于相应条件概率的乘积 $\prod_t P(Y_{it} = y_{it} \mid u_i)$。将这个概率乘积项对服从 $N(0, \hat{\sigma}^2)$ 分布的 u_i 求积分,就得到了所估计的相应单元格的边际概率(即,在不同对象之间求平均)。这里需要用到第 12.6 节将要介绍的数值积分法(numerical integration)。将计算出的边际概率乘以相应的多项分布的样本规模,就得出了该结果组合的拟合值。

表 12.3 关于随机效应模型(式 12.10)的最大似然估计以及相应边际模型的最大似然和 GEE 估计

效 应	参 数	广义线性混合模型的最大似然估计		边际模型的最大似然估计		边际模型的 GEE 估计	
		估计值	SE	估计值	SE	估计值	SE
流产题目	$\beta_1 - \beta_3$	0.83	0.16	0.148	0.030	0.149	0.030
	$\beta_1 - \beta_2$	0.54	0.16	0.098	0.027	0.097	0.028
	$\beta_2 - \beta_3$	0.29	0.16	0.049	0.027	0.052	0.027
性别	γ	0.01	0.48	0.005	0.088	0.003	0.088
$\sqrt{\mathrm{var}(u_i)}$	σ	8.6	0.54				

可以预见的是,对于这些数据,两个性别中回答结果为(0,0,0)和(1,1,1)的拟合值均最大。例如,女性有 440 人对所有三个问题都表示支持(457 人对三个问题都反对),相应的拟合值为 436.5(459.3)。用于比较 16 个观察计数和拟合计数的总体卡方统计量 $G^2 = 23.2$ 和 $X^2 = 27.8$(df = 9)。考虑到样本规模很大,而且我们使用很少的参数($\beta_1, \beta_2, \beta_3, \gamma, \sigma$)来描述表 10.13 中的 14 个多项分布单元格概率(对每个性别有 8 − 1 = 7),这些统计量的值并不算大。这里,因为广义线性混合模型利用五个参数来分析 14 个多项分布参数,所以 df = 9。

通过允许性别与题目之间存在交互效应,可以对上述模型进行扩展。在扩展模型中,男性和女性对应着不同的 $\{\beta_t\}$。但是,该模型并没能显著地改善拟合结果。对扩展模型中额外的参数等于 0 进行检验的似然比统计量为 1.0(df = 2)。

对这些数据的另一种分析则主要关注边际分布的情况,而将数据的相依结构作为次要的。式 12.10 模型所对应的边际模型为

$$\mathrm{Logit}\left[P(Y_t = 1)\right] = \beta_t + \gamma x。$$

对于此模型,表 12.3 也给出了使用可交换的操作性相关结构情况下的 GEE 估计以及最大似然估计的结果。边际模型对这些数据拟合得很好,$G^2 = 1.1$;这里,由于模型利用四个参数描述了六个边际概率(每个性别各三个),所以 df = 2。这些总体平均的 $\{\hat{\beta}_t\}$ 比广义线性

混合模型中的对象别$\{\hat{\beta}_t\}$小得多,这种差别反映了广义线性混合模型中非常大的异质性($\hat{\sigma}=8.6$)以及相应的三个回答之间强相关关系的影响。例如,GEE 分析所估计的每对回答之间的共同相关系数为 0.82。尽管广义线性混合模型的$\{\hat{\beta}_t\}$是边际模型中相应$\{\hat{\beta}_t\}$的五到六倍,二者的标准误也是如此。因而,两种方法所得出的最终结果与实质结论很相似。

12.3.3 对精神抑郁跟踪研究的再讨论

现在,我们重新讨论一下表 11.2 中关于比较新药品与标准药品对治疗精神抑郁的效果的跟踪研究数据。在第 11.2.1 节中,我们利用边际模型分析了这些数据。由 y_t 表示关于精神抑郁的第 t 次测量结果,正常等于 1,不正常则等于 0。根据初诊严重程度 s (1 = 严重,0 = 不严重)、服用药品 d(1 = 新药品,0 = 标准药品),以及测量时点 t,我们构建模型

$$\text{Logit}\left[P(Y_t = 1)\right] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt$$

来评估边际分布的情况。

现在,令 y_{it}表示关于第 i 个对象的第 t 次测量结果。模型

$$\text{Logit}\left[P(Y_{it} = 1) \mid u_i\right] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt + u_i$$

分析的是对象别效应,而不是总体平均效应。表 12.4 给出了该模型的最大似然估计结果。关于标准药品效应的时间趋势的估计为 $\hat{\beta}_3=0.48$,而关于新药品的相应估计为 $\hat{\beta}_3+\hat{\beta}_4=1.50$。这些值与相应的边际模型的最大似然估计和 GEE 估计结果几乎相同,如表 12.4 所示(在第 11.2.1 和第 11.3.2 节中,我们已经对这些结果进行了讨论)。原因正如 GEE 分析所表明的那样,数据中重复测量的观测值之间并没有显示出多大的相关关系。这里,随机效应模型中的 $\hat{\sigma}=0.07$ 也印证了这一点,对象之间基本上不存在异质性。

表 12.4 对表 11.2 数据拟合的边际/条件 Logit 模型的参数估计

参 数	边际模型的最大似然估计	标准误	边际模型的 GEE 估计	标准误	随机效应模型的最大似然估计	标准误
初诊结果	-1.29	0.14	-1.31	0.15	-1.32	0.15
药品	-0.06	0.22	-0.06	0.23	-0.06	0.22
时点	0.48	0.12	0.48	0.12	0.48	0.12
药品×时点	1.01	0.18	1.02	0.19	1.02	0.19

基于模型的拟合结果,对服从 $N(0,0.07^2)$分布的随机效应求积分,可以得到有关结果组合的边际拟合值。将这些拟合值与表 11.2 中的样本计数相对比,可以发现模型拟合得相对较好。该模型利用六个参数描述了 28 个多项分布单元格概率(对初诊严重程度-药品的四种组合来说,每种的三元结果变量分别对应七个)。用来比较单元格的观察计数及其拟合值的常见统计量为 $G^2=22.0$ 和 $X^2=20.8$($\text{df}=28-6=22$)。

当我们假定模型中的 $\sigma=0$ 时,模型的偏离度仅上升了 0.001。根据在第 12.6.6 节将要介绍的结果,进行模型间比较的 P 值等于将偏离度视为 df = 1 的卡方分布所得到的 P 值的一半,这里 $P=0.49$。这个较简单模型所给出的效应估计和标准误估计与原来的结果几乎完全一样,因而它是充分的。通过比较不同模型的 AIC 值也可以得出相同的结论(例如,SAS 的 PROC NLMIXED 给出的随机效应模型的 AIC 等于 1 173.9,而假定 $\sigma=0$

的更简单模型的 AIC 等于 1 171.9)。

12.3.4 利用模型分析多中心临床试验中的异质性

许多应用在两组对象之间比较结果变量的差别时,往往会按照第三个变量进行分层。当在结果变量为二分变量时,这样的数据可以表示为几个 2 ×2 列联表。研究关注的重点是分析这些 2 ×2 表格中的关联,以及在不同层间这种关联是否存在差异。

有时层本身就是一个样本,比如学校或者诊所。这时,随机效应方法是一种自然的选择。在层是随机样本的情况下,随机效应模型可以使我们对这些层所代表的总体进行统计推断。比如,随机效应模型的拟合结果提供了对层的总体的对数发生比之比的估计均值和标准差。在每个层内,它也给出了关于对数发生比之比的一个预测,将样本值向均值进行收缩。当层内样本规模较小,因而普通的样本发生比之比具有较大的标准误时,这一点尤其重要。即使当层本身不是一个随机样本,甚至不是一个样本时,相应的模型仍然具有上面所提到的优点,尽管此时随机效应方法不再是一种自然的选择。

我们通过表 12.5 来加以说明,这些数据取自一项在八个中心开展的临床试验的结果,对此我们曾在第 6.3 节进行过分析。该研究的目的是通过与控制组进行比较来分析一种现行药品治疗某种感染的效果。对于在第 i 个中心使用干预方式 t(1 = 现行药品;2 = 控制)的对象,令 $y_{it}=1$ 表示成功。一种可考虑的模型是 logistic-正态模型,

$$\text{Logit}\left[P(Y_{i1}=1 \mid u_i)\right]=\alpha+\beta/2+u_i,$$

$$\text{Logit}\left[P(Y_{i2}=1 \mid u_i)\right]=\alpha-\beta/2+u_i, \qquad (12.11)$$

其中$\{u_i\}$是独立的 $N(0,\sigma^2)$ 变量。这一模型假定,对八个中心来说,不同干预方式与结果之间的对数发生比之比 β 都相等。参数 σ 描述各中心在成功概率上的异质性。

表 12.5 按干预方式划分的八个中心临床试验的结果

中心	干预方式	结果		样本的发生比之比	拟合的发生比之比
		成功	失败		
1	现行药品	11	25	1.19	2.02
	控制	10	27		
2	现行药品	16	4	1.82	2.09
	控制	22	10		
3	现行药品	14	5	4.80	2.19
	控制	7	12		
4	现行药品	2	14	2.29	2.11
	控制	1	16		
5	现行药品	6	11	∞	2.18
	控制	0	12		
6	现行药品	1	10	∞	2.12
	控制	0	10		
7	现行药品	1	4	2.0	2.11
	控制	1	8		
8	现行药品	4	2	0.33	2.06
	控制	6	1		

来源:Beitler and Landis(1985).

允许干预效果与中心之间存在交互效应的 logistic-正态模型为

$$\text{Logit}\left[P(Y_{i1}=1 \mid u_i, b_i)\right] = \alpha + (\beta + b_i)/2 + u_i,$$

$$\text{Logit}\left[P(Y_{i2}=1 \mid u_i, b_i)\right] = \alpha - (\beta + b_i)/2 + u_i, \tag{12.12}$$

其中$\{u_i\}$是独立的$N(0,\sigma_a^2)$变量，$\{b_i\}$是独立的$N(0,\sigma_b^2)$变量，并且$\{u_i\}$与$\{b_i\}$相互独立。在第 i 个中心，相应的对数发生比之比等于$\beta+b_i$。对数发生比之比在各中心之间按照$N(\beta,\sigma_b^2)$分布变动。也就是说，β是干预方式与结果之间的中心别对数发生比之比的期望值，而σ_b描述了这些对数发生比之比的变动性。模型的参数为$(\alpha,\beta,\sigma_a,\sigma_b)$。

在表 12.5 中，无论是控制组还是干预组，各个中心的样本成功率变动非常大，不过除了最后一个中心外，都是干预组的成功率更高。在使用既包括中心的随机效应又包括干预方式的随机效应的模型时，数据中最好能包括更多的中心。在只有这样少数几个中心的情况下，很难得到关于方差构成的可靠估计。在这里需要注意的是，我们只是通过这些数据来展示一下相应的模型。当数据中包括很多的中心时，模型也可以考虑允许 b_i 和 u_i 之间存在相关关系，但在此我们不考虑这种情况。在不包括交互项的式 12.11 模型中，对干预效应的估计为$\hat{\beta}=0.739$(SE =0.300)；允许交互项的式 12.12 模型的估计结果为$\hat{\beta}=0.746$(SE =0.325)。两个模型的结果都表明，该现行药品具有相当显著的效果。然而，在样本规模这么小的情况下，我们很难断定这个效应到底有多强。

在允许交互项的模型中，支持关联的证据更弱一些。在不包括交互项的模型中，沃尔德统计量为$(0.739/0.300)^2=6.0$；而在包括交互项的模型中，该统计量仅为$(0.746/0.325)^2=5.3$。相应的似然比统计量则分别为 6.3 和 4.6(df =1)。允许交互项的模型中多出的方差项描述了对数发生比之比的变动性。随着该项估计值$\hat{\sigma}_b$的上升，所估计的干预效应$\hat{\beta}$的标准误一般也会上升。在这个例子中，$\hat{\sigma}_b=0.15$相对较小，两个模型中$\hat{\beta}$的标准误相差不大。当$\hat{\sigma}_b=0$时，两个模型的标准误和参数拟合结果完全相同。

为了表明$\hat{\sigma}_b$值较大时对所估计的平均干预效应$\hat{\beta}$的标准误的影响，我们对表 12.5 的数据进行了微小的改动：将第 3 个中心的干预组中的三个失败结果改为成功，将第 8 个中心的干预组中的三个成功结果改为失败。在对数据进行改动后，所估计的干预效应的变动性从$\hat{\sigma}_b=0.15$上升到$\hat{\sigma}_b=1.4$。这时，关于平均干预效应的估计在不包括交互项的式 12.11 模型中为$\hat{\beta}=0.722$(SE = 0.299)，而在包括交互项的模型中为$\hat{\beta}=0.767$(SE =0.623)。相应的沃尔德统计量分别为 5.8 和 1.5。因而，在包括交互项的式 12.12 模型中，干预效应变得相当微弱。这并不奇怪，当干预效应在各中心之间变化很大时，对其平均值的估计也就变得更加困难。

回到表 12.5 的实际数据，由于在式 12.12 模型中$\hat{\sigma}_b=0.15$相对很小，该模型对样本发生比之比进行了明显的收缩。表 12.5 给出了发生比之比的样本值以及模型的拟合值。这些结果是基于对随机效应的预测(我们将在第 12.6 节予以解释)，并将预测结果以及关于固定效应的最大似然估计代入模型公式中，分别计算每个中心在每种干预方式下两种结果出现的概率。样本的发生比之比的变动范围是 0.33 到∞，而相应的随机效应模型的拟合值(由 SAS 中的 PROC NLMIXED 估计)的变动范围仅为 2.0 到 2.2。经过模型修匀后，估计值的变动性下降了很多，而且各中心的排序与样本值也并不完全一致。例如，对第 3 个中心的修匀估计为 2.2，大于对第 6 个中心的修匀估计 2.1，尽管后者的样

本值是无穷大。这一结果的部分原因是由于样本规模较小,因而收缩的幅度较大。当 $\hat{\sigma}_b=0$ 时,式 12.12 模型给出的拟合结果与式 12.11 模型完全相同,并且对每个中心所估计的发生比之比也都相等。

有关多个 2×2 表格间的异质性的研究,参见:Liu 和 Pierce(1993)、Skene 和 Wakefield(1990)。

12.3.5 随机效应模型的其他表述形式

随机效应模型还可以通过其他方法来表示。例如,包括交互项的式 12.12 模型可以等价地表示为

$$\text{Logit}\left[P(Y_{it}=1\mid u_i,b_{it})\right]=\alpha+\beta x_t+b_{it}+u_i,$$

其中 x_t 是代表干预方式的虚拟变量($x_1=1,x_2=0$),$\{u_i\}$ 是独立的 $N(0,\sigma_a^2)$ 变量,$\{b_{i1}\}$ 和 $\{b_{i2}\}$ 都是独立的 $N(0,\sigma^2)$ 变量。这里,$b_{i1}-b_{i2}$ 对应着式 12.12 中的参数 b_i,而 $2\sigma^2$ 则对应着 σ_b^2。

在构建随机效应模型时,需要特别留意模型表达式的含义以及随机效应之间的相关结构。假如我们将包括交互项的式 12.12 模型表示为

$$\text{Logit}\left[P(Y_{it}=1\mid u_i,b_i)\right]=\alpha+(\beta+b_i)x_t+u_i,\qquad(12.13)$$

其中 $\{b_i\}$ 服从 $N(0,\sigma_b^2)$ 分布。这种表述是不恰当的,因为当 $x_2=0$ 时,$\{u_i\}$ 和 $\{b_i\}$ 不存在相关关系,而该模型所设定的第一种干预方式的 Logit 的变动性比第二种大。同时,模型也不应当依赖于虚拟变量 x_t 的定义。然而,值得注意的是,如果对某个常数 c 有 $z_t=x_t+c$,那么式 12.13 模型等价于

$$\begin{aligned}&\text{Logit}\left[P(Y_{it}=1\mid u_i,b_i)\right]\\&=\alpha+(\beta+b_i)(z_i-c)+u_i=\alpha'+(\beta+b_i)z_i+v_i,\end{aligned}$$

其中 $\alpha'=\alpha-c\beta,v_i=u_i-cb_i$。因此,即便($u_i,b_i$)不相关,($v_i,b_i$)仍存在相关关系。实际上,只有当随机效应之间存在相关关系时,表达式 12.13 才适用。这时,它等价于在式 12.12 模型中允许随机效应之间相关的情形。有关进一步的讨论,参见:Agresti 和 Hartzel(2000)。

12.3.6 用来预测总体规模的捕获-再捕获模型

捕获-再捕获实验是一种利用一系列样本来估计总体规模的方法。传统上,这种方法被广泛应用于估算某个栖息地的动物数量。在每次进行抽样时,对被捕获的动物打上某种标记。将已捕获的动物全部放生,并将所有动物作为下一次抽样的对象总体。在进行 T 次抽样后,可以用一个 2^T 列联表表示有关数据,其中每次的结果为(被捕获,未被捕获)。每次均对应着未被捕获的单元格计数 $n_{22\ldots2}$ 是缺失的。如果我们知道这个计数,将其与其他单元格计数相加就会得出总体规模。可以对这个 2^T 表格设定模型,并利用 2^T-1 个观察到的计数来拟合它。模型拟合过程对应的是 2^T-1 个观察到的单元格,但是通过对模型进行外推,我们可以得到有关未观察到的单元格计数的估计。将该估计值与 2^T-1 个观察到的计数加总,就得到了关于总体规模的估计。

举例来说,假定 $T=2$。我们在两次抽样中都观察到的动物数为 n_{11},只在第一次观察到而在第二次未观察到的动物数是 n_{12},只在第二次观察到而在第一次未观察到的动物数是

n_{21}。我们并不知道两次都没有被捕获的动物数量 n_{22}。如果假定这个 2×2 表格满足独立性，预测值 $\hat{n}_{22}$ 应当是使表格的发生比之比等于 1.0 的数；根据 $(n_{11}\hat{n}_{22})/(n_{12}n_{21})=1$，可推出 $\hat{n}_{22}=n_{12}n_{21}/n_{11}$。由此得出，关于总体规模的预测（Sekar and Deming，1949）为

$$\hat{N} = n_{11} + n_{12} + n_{21} + n_{12}n_{21}/n_{11} = n_{1+}n_{+1}/n_{11},$$

并且 $\widehat{\text{var}}(\hat{N}) = \dfrac{n_{1+}n_{+1}n_{12}n_{21}}{n_{11}^3}$。然而，独立性假设往往并不现实。当具有多于两次的抽样信息时，我们可以尝试使用更复杂的模型。

表 12.6 取自一项 $T=6$ 天的连续捕获雪兔（snowshoe hare）以估计其总体规模的研究，Cormack（1989）及其他研究者曾经分析过该数据。该研究共观察到 68 只雪兔。如表 12.6 所示，在第一天共观察到 3 只雪兔，但后面几天都没有再观察到这几只。为简便起见，在短期研究中，我们假定这个总体没有发生死亡、出生或迁移，称之为一个*封闭总体*（*closed population*）。

表 12.6 雪兔的捕获-再捕获结果

捕获 6	捕获 5	捕获 4	捕获 3，捕获 2，捕获 1[a] 000	001	010	011	100	101	110	111
0	0	0	—	3	6	0	5	1	0	0
			(24.0)	(2.3)	(5.4)	(0.9)	(3.2)	(0.5)	(1.2)	(0.3)
0	0	1	3	2	3	0	0	1	0	0
			(4.8)	(0.8)	(1.8)	(0.5)	(1.1)	(0.3)	(0.6)	(0.3)
0	1	0	4	2	3	1	0	1	0	0
			(3.9)	(0.6)	(1.5)	(0.4)	(0.9)	(0.2)	(0.5)	(0.2)
0	1	1	1	0	0	0	0	0	0	0
			(1.3)	(0.3)	(0.8)	(0.3)	(0.5)	(0.2)	(0.4)	(0.3)
1	0	0	4	1	1	1	2	0	2	0
			(6.8)	(1.1)	(2.6)	(0.6)	(1.5)	(0.4)	(0.9)	(0.4)
1	0	1	4	0	3	0	1	0	2	0
			(2.3)	(0.6)	(1.3)	(0.5)	(0.8)	(0.3)	(0.7)	(0.4)
1	1	0	2	0	1	0	1	0	1	0
			(1.9)	(0.5)	(1.1)	(0.4)	(0.7)	(0.3)	(0.6)	(0.4)
1	1	1	1	1	1	0	0	0	1	2
			(1.0)	(0.4)	(0.9)	(0.5)	(0.5)	(0.3)	(0.7)	(0.7)

a 括号中的数值表示 logistic 正态模型的拟合值。

来源：A. Agresti，*Biometrics* 50：494-500（1994）。

大多数关于捕获-再捕获的方法都假设在给定时点每个对象（即动物）具有相同的被捕获概率。通常这种假设是不现实的。允许被捕获概率存在异质性的一种办法是拟合包括对象层面的随机效应的 Logit 模型。对于第 i 个对象，$i=1,\cdots,N$ 且 N 未知，令 $y'_i=(y_{i1},\cdots,y_{iT})$，其中 $y_{it}=1$ 表示在第 t 个样本中被捕获，$y_{it}=0$ 则表示未被捕获。在缺乏解释变量的情况下，我们可以考虑使用 Rasch 形式的模型

$$\text{Logit}\left[P(Y_{it}=1 \mid u_i)\right] = u_i + \beta_t,$$

其中 $\{u_i\}$ 服从独立的 $N(0,\sigma^2)$ 分布。β_t 的值越大，在第 t 次抽样中被捕获的概率也就越大。σ 的值越大，被捕获概率的异质性也就越强。当 $\sigma=0$ 时，这个 logistic-正态模型简化

为关于 2^T 表格的相互独立性模型(即式 8.6 对数线性模型)。

与其他随机效应模型相同,根据($\mathbf{y}_i|u_i$)的概率密度函数对随机效应求积分得出似然函数(这一点将在第 12.6 节讨论)。对在未观察到的单元格中所有可能的计数,考虑这个似然函数以及相应的关于$\{\beta_t\}$和 σ 的最大似然估计。剖面似然函数(profile likelihood function)将最大似然函数视为关于未观察到的单元格计数的一个函数。使该剖面似然函数最大化的值,就是对未观察到的单元格计数的最大似然预测。在缺乏专门软件的情况下,我们可以通过试错法——赋予未观察到的单元格计数各种不同的值来不断拟合随机效应模型——以求出使剖面似然函数最大化的计数。

通过最大似然法对表 12.6 的数据拟合这个模型,得出对未观察到的单元格计数的预测值为 24。由于该研究观察到了 68 只雪兔,所以关于雪兔的总体规模的估计为 $\hat{N}=92$。在模型的拟合结果中,$\hat{\sigma}=1.0$。

利用剖面似然函数法或者非参数自举重抽样法(bootstrap)可以获取关于 N 的置信区间。在剖面似然函数法中,关于缺失的单元格计数的区间包括满足下列条件的可能计数:相应的 G^2 拟合统计量与最大似然估计时的取值相比上升幅度小于$\chi_1^2(\alpha)$。将样本中被观察到的对象与这个区间的端点相加,就得到了关于 N 的相应区间。在雪兔的例子中,关于 N 的 95% 的剖面似然置信区间是(75,154)。$\hat{N}$ 与区间的下限更为接近,这样的情况很常见。有关细节,可参见:Coull 和 Agresti(1999)。

异质性程度越高,即 $\hat{\sigma}$ 取值越大,一般 $\hat{N}$ 也会比较大,而且相应的置信区间也越宽。$\hat{\sigma}$ 的值很大会增加估计的难度,因为它会导致似然函数的平面相对扁平,这就意味着对 N 的估计更不精确。需要特别指出的是,当似然函数足够扁平时,关于 N 的剖面似然置信区间的上限等于无穷大。另外,这时的最大似然估计值往往不稳定,数据的微小变动会导致 $\hat{N}$ 的很大变动。当被捕获的概率很小时,也会出现类似的问题。当大多数被捕获的对象都只出现在某一个样本中时,就表明存在这种情况。在出现上述情况或者当 $\hat{\sigma}$ 很大时,希望 N 的置信区间很小是不现实的。

有关的其他模型,我们将在第 13.1.3 节讨论。当忽略了可能的异质性时,模型所得到的关于 N 的置信区间往往严重偏小。尽管在传统上捕获-再捕获方法主要应用于对动物总体的研究,这种方法同样也适用于估计人类总体,比如估计人口中注射吸毒或者感染 HIV 的比例。Darroch 等(1993)用此方法考察了有关普查人口的估计,Chao 等(2001)估计了在一次肝炎爆发期内被感染的人口数(习题 12.21)。另一个关于该方法的有趣应用是,利用几个搜索引擎抽取样本,以估计万维网(World Wide Web)中与某一主题有关的文件数量(Fienberg et al., 1999)。

12.4 多项分布数据的随机效应模型

关于二分结果变量的随机效应模型可以扩展到多类别结果变量的情况。在第 7 章所介绍的多类别结果变量的模型中,对于具有 I 个类别的多项分布观测值,相当于在模型中包括一个具有 $I-1$ 个指示变量的向量,当观测值落入第 j 个类别时,第 j 个指示变量等于 1,否则为 0。在第 7.1.5 节中,通过对多元结果变量应用一个由连结函数组成的向量,我们定义了多元广义线性模型。在多元广义线性模型中加入随机效应,将其以及广义线性混合模型

式 12.1 扩展为多元广义线性混合模型(Hartzel et al., 2001b;Tutz and Hennevogl, 1996)。它既包括有关定类结果变量的模型,也包括有关定序结果变量的模型。

12.4.1 包括随机截距的累积 Logit 模型

定序结果变量的模型分析往往比对定类结果变量更容易一些,因为在前者的情况下,通常可以对每个 Logit 设定相同的随机效应和固定效应。对于累积 Logit 模型来说,这相当于*比例发生比*(*proportional odds*)的结构(第 7.2.2 节)。将第 i 个群组的第 t 个观测值的结果变量表示为 y_{it},它的可能取值为 $1,2,\cdots,I$。关于累积 Logit 的广义线性混合模型可以表示为:

$$\text{Logit}\left[P(Y_{it} \leqslant j \mid \mathbf{u}_i)\right] = \alpha_j + \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i, j = 1,\cdots,I-1。\qquad (12.14)$$

Hedeker 和 Gibbons(1994)讨论了有关的模型拟合方法,主要是将 $\mathbf{u}_i$ 视为服从多元正态分布。

与第 12.2.2 节介绍的有关二分结果变量的情况相同,对于带有随机截距的累积 Logit/probit 模型,随机效应模型中的参数也和边际模型的相应参数存在类似的关系。边际效应的值一般小于随机效应的值,并且随着 σ 的增加,二者之间的差异也会变大。同时,当其他连结函数允许对每个 Logit 都具有相同效应时,式 12.14 模型的预测项结构仍然成立。例如,Hartzel 等(2001a,b)将此预测项结构用于对相邻类别 Logit 的分析。

12.4.2 关于失眠症研究的再讨论

表 11.4 取自一项临床试验的结果,该研究在两个时点比较一种现行药品与安慰剂治疗失眠症的效果。在第 11.2.3 节和第 11.3.3 节中,我们利用边际模型对该数据进行了分析。令 y_t = 在时点 t 的入睡时间,边际模型

$$\text{Logit}\left[P(Y_t \leqslant j)\right] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 tx$$

允许 t = 时点(0 = 初期,1 = 跟踪期)和 x = 干预方式(1 = 现行药,0 = 安慰剂)之间存在交互效应。表 12.7 给出了关于该模型的最大似然估计和 GEE 估计的结果。

现在,令 y_{it} 表示第 i 个对象在时点 t 的结果变量。表 12.7 还给出了随机截距模型

$$\text{Logit}\left[P(Y_{it} \leqslant j \mid u_i)\right] = u_i + \alpha_j + \beta_1 t + \beta_2 x + \beta_3 tx$$

的拟合结果。随机效应模型的实质性结论与边际模型很相似,但是,前者的参数估计和标准误都比后者大了大约 50%。这反映了数据中存在着较大的异质性,以及由此导致的两个时点的结果变量之间的强关联。

表 12.7 对表 11.4 数据拟合的累积 Logit 模型[a]

效 应	边际模型的最大似然估计	边际模型的GEE 估计	随机效应模型(广义线性混合模型)的最大似然估计
干预方式	0.046(0.236)	0.034(0.238)	0.058(0.366)
时点	1.074(0.162)	1.038(0.168)	1.602(0.283)
干预方式 × 时点	0.662(0.244)	0.708(0.244)	1.081(0.380)

a 括号中的数值为标准误。

12.4.3 整群抽样

在分析通过整群抽样得到的调查数据时,需要对基于简单随机样本(如单一的多项

分布样本)的标准分析方法进行一定的调整。普通方法给出的标准误太小;常规的卡方检验统计量不再服从卡方零分布,而是服从多个卡方的加权求和。针对分类数据分析和建模,Rao 和 Thomas(1988)综述了在有关分析中将复杂抽样设计考虑进来以对标准的统计推断进行调整的方法。

在抽样设计中,如果群组的抽取是随机的,我们可以利用群组层面的随机效应来处理相应的群组问题。我们以 Brier(1980)的数据为例来加以展示,该数据包括来自 20 个社区(群组)的 96 个观测值,其中 Y = 家庭满意度,X = 整个社区的满意度。每个变量的测量尺度都是定序尺度(不满意,满意,非常满意)。在 Brier 的分析中,当调整了群组效应后,对由 X 和 Y 组成的 3 × 3 列联表进行独立性检验的皮尔逊统计量从 17.9 下降到 15.7(df = 4)。

针对第 i 个群组的第 t 个观测值 y_{it},考虑模型

$$\text{Logit}\left[P(Y_{it} \leqslant j \mid u_i)\right] = u_i + \alpha_j + x_{it}\beta, \tag{12.15}$$

其中,关于 x_{it} 的满意水平的赋值为(1,2,3)。假定 u_i 服从 $N(0,\sigma^2)$ 分布,最大似然估计结果为 $\hat{\beta} = -1.201(\text{SE} = 0.407)$,并且 $\hat{\sigma} = 0.92$。相反,将 96 个观测值作为一个随机样本相当于在拟合上述模型时令 $\sigma = 0$。这时的拟合结果为 $\hat{\beta} = -1.226(\text{SE} = 0.370)$。在调整群组问题后,显著性水平出现了微小的下降。

12.4.4　包括随机效应的基线类别 Logit 模型

对于定类结果变量,我们可以分别构建一个基线类别与其他每个类别之间的二分模型,同时拟合这些模型,并允许每个模型具有不同的效应。这需要指定一个由群组别随机效应组成的向量 $\mathbf{u}_{ij}$,其中每个 Logit 各对应一个。包括随机效应的基线类别 Logit 模型的一般形式可表示为:

$$\log\frac{P(Y_{it} = j)}{P(Y_{it} = I)} = \alpha_j + \mathbf{x}'_{it}\boldsymbol{\beta}_j + \mathbf{z}'_{it}\mathbf{u}_{ij}, \quad j = 1,\cdots,I-1。$$

这里,因为基线类别的选取是随意的,固定效应 $\boldsymbol{\beta}_j$ 和随机效应 $\mathbf{u}_{ij}$ 都取决于 j。在结果变量为定类变量的情况下,没有理由认为不同的 j 所对应的效应会相似。

第 i 个群组的随机效应向量为 $\mathbf{u}'_i = (\mathbf{u}'_{i1},\ldots,\mathbf{u}'_{i,I-1})$。通常的方法是将 $\{\mathbf{u}_i\}$ 视为独立的多元正态变量,我们则建议对 $\mathbf{u}_i$ 使用一个未设定的协方差矩阵 $\boldsymbol{\Sigma}$。例如,允许不同 Logit 所对应的随机效应具有不同的方差是很合理的。在设定同方差的情况下,这个方差不会与由任意两个类别的 Logit(如 $\log[P(Y_{it} = j)/P(Y_{it} = k)]$)所具有的随机效应的方差都一样。在使用未设定的协方差时,无论选取哪个类别作为基线类别,模型的结构都相同。这样的例子,可参见:Hartzel 等(2001b)。

12.5　二分数据的多元随机效应模型

在现实应用中,随机效应常常是单维的,表现为随机截距的形式。然而,我们已经看到,关于定类结果变量的模型要求多元随机效应;另外,二元随机效应对于描述多中心临床试验中的异质性很有帮助。在本节中,我们介绍一些其他多元随机效应模型的例子。

12.5.1　二元二分结果变量的配对数据

Leo Goodman 在多篇文章里分析了表 12.8 的数据(如 Goodman, 1974)。该数据中,

一个学校男生的样本先后被访问了两次,中间相隔几个月。他们被问及是否认同自己属于“前卫群体(leading crowd)”中的一员,以及隶属于这一群体是否有时需要违背自己的原则。因此,这里包括两个二分结果变量,我们分别称之为“成员认同”和“态度”,同时对每个对象都进行了两次调查。表 12.8 将态度的类别标示为(正面,负面),其中“正面”代表不同意必须违背自己的原则这一论断。

表 12.8 关于“前卫群体”的成员认同与态度

第一次调查时的(M,A)	第二次调查时的(M,A)[a]			
	(属于,正面)	(属于,负面)	(不属于,正面)	(不属性,负面)
属于,正面	458	140	110	49
属于,负面	171	182	56	87
不属于,正面	184	75	531	281
不属于,负面	85	97	338	554

a M:成员认同;A:态度。

来源:J. S. Coleman, *Introduction to Mothematical Sociology* (London: Free Press of Glencoe, 1964), p. 170。

对于第 i 的对象,令 y_{itv} 表示其在第 t 个调查时点关于变量 v 的回答结果,其中 $v = M$ 代表成员认同,$v = A$ 代表态度。Logit 模型

$$\text{Logit}\left[P(Y_{itv} = 1 \mid u_{iv})\right] = \beta_{tv} + u_{iv} \tag{12.16}$$

相当于一个多元形式的 Rasch 模型(式 12.4)。它包括每个变量 v 的可加效应和对象别效应。这里,(u_{iM}, u_{iA})是二元随机效应,用来描述(成员认同,态度)的对象间异质性。我们假定$\{(u_{iM}, u_{iA})\}$服从独立的二元正态分布 $N(\mathbf{0}, \mathbf{\Sigma})$,二者的方差不等,并且存在非零的相关关系。

对该模型进行最大似然拟合,得到 $\hat{\beta}_{2M} - \hat{\beta}_{1M} = 0.379$(SE = 0.075)以及 $\hat{\beta}_{2A} - \hat{\beta}_{1A} = 0.176$(SE = 0.058)。对这两个变量来说,在第二次访谈中回答结果落在第一个类别的概率都更高一些。例如,对于一个给定对象,在第二次访谈中所估计的自我认定属于前卫群体成员的发生比是第一次访谈时相应发生比的 exp(0.379) = 1.46 倍。

所估计的两个随机效应之间的相关系数等于 0.30,$\{u_{iM}\}$的标准差为 $\hat{\sigma}_1 = 3.1$,$\{u_{iA}\}$的标准差为 $\hat{\sigma}_2 = 1.5$。由于二者的标准差相差较大,在边际模型和条件模型中,有关成员认同和态度的效应的相对大小有所不同(回忆我们在第 12.2.3 节所做的说明)。对成员认同来说,边际效应收缩的程度更大。具体而言,在这个条件模型中,两个变量的发生比之比估计值的比率等于 exp(0.379)/exp(0.176) = 1.46/1.19 = 1.22;而在边际模型中,可以通过每个变量在每个时点的边际分布来估计发生比之比(例如,对于成员认同,其值等于(1 392/2 006)/(1 253/2 145) = 1.188),相应的两变量的发生比之比估计值的比率仅为 1.188/1.133 = 1.05。

对所估计的随机效应分布求积分,就得出了表 12.8 中的 16 种可能取值组合的拟合值。将 16 个观察计数与其拟合值相比较,模型的偏离度 $G^2 = 5.5$(df = 8)。这个模型通过七个参数描述了 15 个多项分布概率,它对数据拟合得很好。相比之下,限定随机效应之间不存在相关关系的模型则拟合得很差($G^2 = 97.5$, df = 9)。限定随机效应之间存在完全相关的模型则等价于每个对象只具有单一的随机效应 u_i。这时,模型相当于一个包括四个项目的 Rasch 模型,这四个项目由观测时点与变量之间的组合构

成。该模型对数据拟合得非常糟糕($G^2=655.5$, df = 10)。Agresti 等(2000)对此进行了更详细的讨论。

12.5.2 群组定序结果变量的递进比 Logit:毒性研究

在结果变量为定序变量的递进比 Logit 模型中,这些 Logit 对应的是独立的二项分布变量(第 7.4.3 节)。因而,可以通过二分的 Logit 随机效应模型来分析有关群组定序结果变量的递进比 Logit(Ten Have and Uttal, 1994)。对于第 i 个群组中的第 t 个观测值,令 $\omega_{ij}=P(Y_{it}=j\mid Y_{it}\geqslant j,u_{ij})$(在更一般的情况下,这个概率同时还取决于 t,但是对下面的例子而言,我们暂不需要这种扩展)。相应的递进比 Logit 为$\{\text{Logit}(\omega_{ij}),j=1,\cdots,I-1\}$。

令 n_{ij} 表示在群组 i 中选择类别 j 的对象的数量,$n_i=\sum_{j=1}^{I}n_{ij}$。在递进比 Logit 模型中,对于一个给定的群组,将$(n_{i1},\cdots,n_{i,I-1})$视为多项分布变量等价于将其当作一个独立的二项分布变量的序列集,其中 n_{ij} 是一个 $\text{bin}(n_i-\sum_{h<j}n_{ih},\omega_{ij})$ 变量,$j=1,\cdots,I-1$。

我们通过一项由美国国家毒理学项目中心(the U.S. National Toxicology Program)所开展的关于发育毒性的研究来加以说明。该研究给怀孕母鼠分别服用四种剂量(0, 0.75, 1.50, 3.00 克/千克)的乙二醇(ethylene glycol, EG),以检验其对发育的影响。每组所包括的母鼠数量分别为(25, 24, 22, 23)。在这里,群组对应于一次分娩的一窝子鼠。对每个胎儿而言,三种可能的结果(死亡/消溶,畸形,正常)形成一个定序变量,其中“正常”是最好的结果。表 12.9 给出了有关数据。在这个例子中,由于各类别之间存在层级关系,在一只动物出现畸形之前,它必须得存活,因而递进比 Logit 是很自然的选择。下面的分析取自:Coull 和 Agresti(2000)。

表 12.9 有关 94 窝子鼠的状况分布(死亡数量,畸形数量,正常数量)

剂量 = 0.00 g/kg	剂量 = 0.75 g/kg	剂量 = 1.50 g/kg	剂量 = 3.00 g/kg
(1,0,7), (0,0,14)	(0,3,7), (1,3,11)	(0,8,2), (0,6,5)	(0,4,3), (1,9,1)
(0,0,13), (0,0,10)	(0,2,9), (0,0,12)	(0,5,7), (0,11,2)	(0,4,8), (1,11,0)
(0,1,15), (1,0,14)	(0,1,11), (0,3,10)	(1,6,3), (0,7,6)	(0,7,3), (0,9,1)
(1,0,10), (0,0,12)	(0,0,15), (0,0,11)	(0,0,1), (0,3,8)	(0,3,1), (0,7,0)
(0,0,11), (0,0,8)	(2,0,8), (0,1,10)	(0,8,3), (0,2,12)	(0,1,3), (0,12,0)
(1,0,6), (0,0,15)	(0,0,10), (0,1,13)	(0,1,12), (0,10,5)	(2,12,0), (0,11,3)
(0,0,12), (0,0,12)	(0,1,9), (0,0,14)	(0,5,6), (0,1,11)	(0,5,6), (0,4,8)
(0,0,13), (0,0,10)	(1,1,11), (0,1,9)	(0,3,10), (0,0,13)	(0,5,7), (2,3,9)
(0,0,10), (1,0,11)	(0,1,10), (0,0,15)	(0,6,1), (0,2,6)	(0,9,1), (0,0,9)
(0,0,12), (0,0,13)	(0,0,15), (0,3,10)	(0,1,2), (0,0,7)	(0,5,4), (0,2,5)
(1,0,14), (0,0,13)	(0,2,5), (0,1,11)	(0,4,6), (0,0,12)	(1,3,9), (0,2,5)
(0,0,13), (1,0,14)	(0,1,6), (1,1,8)		(0,1,11)
(0,0,14)			

来源:取自:C. J. Price, C. A. Kimmel, R. W. Tyl, and M. C. Marr, *Toxicol. Appl. Pharmacol.* 81:113-127(1985).

对于第 d 个剂量组的第 i 窝子鼠,令 $\text{Logit}(\omega_{i(d)1})$ 表示死亡概率的递进比 Logit, $\text{Logit}(\omega_{i(d)2})$ 表示在给定存活情况下出现畸形的条件概率的递进比 Logit(下标 $i(d)$ 代表嵌套在剂量组 d 内的第 i 窝子鼠)。令 x_d 表示第 d 组的剂量水平。我们通过窝别随机效应 $\mathbf{u}_{i(d)}=(u_{i(d)1},u_{i(d)2})$ 来反映群组效应,其中 $\mathbf{u}_{i(d)}$ 服从 $N(\mathbf{0},\boldsymbol{\Sigma}_d)$ 分布。这个二元随机效应允许每窝子鼠的死亡概率以及在存活情况下发生畸形的条件概率具有不同程度的过度离散。如果还允许对不同递进比 Logit 具有不同的固定效应,可以考虑模型

$$\text{Logit}(\omega_{i(d)j}) = u_{i(d)j} + \alpha_j + \beta_j x_d。 \quad (12.17)$$

表 12.10 给出了拟合上述模型的四个特例所导致的最大对数似然值的变化情况：

1. 两个 Logit 具有相同的截距和斜率：$\alpha_1 = \alpha_2$ 且 $\beta_1 = \beta_2$；
2. 四个剂量组具有相同的协方差矩阵：$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3 = \mathbf{\Sigma}_4$；
3. 协方差矩阵相同并且随机效应之间不存在相关关系；
4. 不同剂量组之间具有相同的单维方差构成：$u_{i(d)1} = u_{i(d)2}$ 且 $\sigma_d = \sigma$。

表 12.10　关于发育毒性研究的不同多元随机效应模型的对数似然值

模　型	参数数量	参数数量的变动	对数似然值的变动
剂量别 Σ_i	16	—	—
Σ_i，相同的 α,β	14	2	28.4
相同的 Σ	7	9	7.4
相同的 Σ，$\rho = 0$	6	10	7.4
单维 σ^2	5	11	16.7

通过普通的似然比检验可以对前三个特例与一般模型（式 12.17）进行比较。使用具有相同协方差结构并且随机效应之间不相关的较简单模型（即表 12.10 中的第四个模型）基本不会损失任何信息，将其与式 12.17 模型相对比的似然比统计量等于 2(7.4) = 14.8(df = 10)。该模型对每个条件二项分布结果分别应用一个单维随机效应 logistic-正态模型，并设定不论是在同一窝内还是就边际分布而言，死亡子鼠所占比例与（给定存活）畸形子鼠所占比例是独立的。

表 12.10 中的单维模型是所列出的第三个模型的特例，相当于在第三个模型中假定两个 Logit 具有同方差并且随机效应之间存在完全相关，因而，它简化为一个单维随机效应模型。将单维随机效应模型与相应的多元模型进行比较，需要检验关于相关系数的参数是否落在边界上。只有当参数落在参数空间内部的时候，似然比检验的普通卡方渐近理论才适用。对于零模型（null model）的相关系数取值为 1 或者方差构成为 0 的情况，进行相应的检验非常复杂，超出了这里讨论的范围（参见第 12.6.6 节）。但是，关于对数似然值变化的简略分析表明，单维随机效应模型在这里拟合得不够充分。

对每个条件二项分布结果分别应用单维随机效应 logistic-正态模型的最大似然估计结果为 $\hat{\beta}_1 = 0.08$（SE = 0.21），$\hat{\beta}_2 = 1.79$（SE = 0.22）。在一个给定的群组中，没有证据表明剂量水平对死亡率有影响，但是对于给定存活情况下所估计的出现畸形的发生比来说，每增加 1 克/千克（g/kg）剂量的乙二醇会导致其上升 exp(1.79) = 6.0 倍。关于方差构成的估计显示，给定存活条件下发生畸形的窝别效应（$\hat{\sigma}_2 = 1.6$）比发生死亡的窝别效应（$\hat{\sigma}_1 = 0.5$）更大。

12.5.3　层级（多层次）模型分析

在教育研究中，层级数据非常普遍，即在不同层次上都存在包括多个分析单位的群组现象。一项关于学生表现影响因素的全国性研究可能会考察学生在一套测试中的成绩，但在数据分析时，使用一个考虑到学生本人、学校或校区，以及县级单位等不同层面的影响的模型。正如对同一个学生进行两次观测比对不同学生进行的观测结果更相似一样，来自同一所学校的两个学生可能也会比来自不同学校的学生更相似。因而，可以将学生、学校，以及县级单位层面的项视为随机效应，并且不同的项对应着模型的*不同层*

次(*levels*)。例如,在模型中,可以将学生作为第 1 层,学校作为第 2 层,以及县作为第 3 层。分析这种具有层级群组关系的数据的广义线性混合模型被称为*多层次模型*(*multilevel models*)。在多层次模型中,可以在模型的每个层级水平上加入随机效应。

我们通过一个二层模型来加以说明。令 $\pi_{i(j)t}$ 表示第 j 所学校的第 i 个学生在一套测试中通过第 t 项测试的概率。包括学生层面和学校层面的随机效应,以及相应解释变量的固定效应的多层次模型可以表示为

$$\text{Logit}\left[\pi_{i(j)t}\right] = \mathbf{x}'_{i(j)t}\boldsymbol{\beta} + u_j + v_{i(j)}\text{。}$$

这里,解释变量 $\mathbf{x}$ 可以包括该套测试中每项测试的特征变量。学校层面的随机效应 u_j 与学校内部学生层面的随机效应 $v_{i(j)}$ 相互独立,并具有不同的方差构成。第 1 层的随机效应 $\{v_{i(j)}\}$ 描述不同学生之间在个人能力、家庭社会经济地位或其他未被 $\mathbf{x}$ 所测量的特征方面的变动性。当 $\{v_{i(j)}\}$ 具有较大的方差构成时,同一学生所进行的不同测试结果之间就存在很强的关联。第 2 层的随机效应 $\{u_j\}$ 描述由诸如学校预算生均支出等未测量的变量所导致的学校间的变动性。

有关包括多元随机效应的多层次模型的例子,可参见:Aitkin 等(1981)、Anderson 和 Aitkin(1985)、Gibbons 和 Hedeker(1997)、Goldstein(1995)、Goldstein 和 Rasbash(1996)、Longford(1993)。

12.6 广义线性混合模型的拟合、推断与预测

广义线性混合模型的拟合是相当复杂的,其主要困难在于,似然方程不存在封闭形式的解。对于存在多元随机效应的模型来说,利用数值法对似然函数进行近似的运算量非常大。在本节中,我们介绍通过最大似然法拟合广义线性混合模型的基本思路,其中的一些方法可以通过有关软件(如 SAS 的 PROC NLMIXED)来应用。

12.6.1 边际似然函数与最大似然拟合

广义线性混合模型是一个二阶段(two-stage)模型。在第一阶段,以随机效应为条件,假定观测值满足一个广义线性模型。也就是说,第 i 个群组中的观测值 y_{it} 服从一个指数族分布,其期望值 μ_{it} 与线性预测项相连结,

$$g(\mu_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i\text{。}$$

这里,$\mathbf{z}'_{it}\mathbf{u}_i$ 是一个已知的抵消项(offset),并且群组内的观测值相互独立。在模型的第二阶段,假定随机效应 $\{\mathbf{u}_i\}$ 服从独立的 $N(\mathbf{0},\boldsymbol{\Sigma})$ 分布。

对于一个离散变量,令 $\mathbf{y}$ 表示由所有观测值所形成的向量,$\mathbf{u}$ 表示由所有随机效应所形成的向量。令 $f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})$ 表示在给定 $\mathbf{u}$ 后 $\mathbf{y}$ 的条件密度函数,$f(\mathbf{u};\boldsymbol{\Sigma})$ 表示关于 $\mathbf{u}$ 的正态密度函数。广义线性混合模型的似然函数 $\ell(\boldsymbol{\beta},\boldsymbol{\Sigma};\mathbf{y})$ 就是 $\mathbf{y}$ 的概率密度函数 $f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\Sigma})$,将其视为关于 $\boldsymbol{\beta}$ 和 $\boldsymbol{\Sigma}$ 的函数。它是对随机效应求积分后有关 $\mathbf{y}$ 的边际分布,

$$\ell(\boldsymbol{\beta},\boldsymbol{\Sigma};\mathbf{y}) = f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\Sigma}) = \int f(\mathbf{y}\mid\mathbf{u};\boldsymbol{\beta})f(\mathbf{u};\boldsymbol{\Sigma})d\mathbf{u}\text{。} \qquad (12.18)$$

该函数通常被称为*边际似然函数*(*marginal likelihood*)。例如,关于 logistic-正态模型(式 12.5)的似然函数 $\ell(\boldsymbol{\beta},\sigma^2;\mathbf{y})$(将 α 包括在 $\boldsymbol{\beta}$ 中)为

$$\prod_i\left(\int_{-\infty}^{\infty}\prod_t\left[\frac{\exp(\mathbf{x}'_{it}\boldsymbol{\beta}+u_i)}{1+\exp(\mathbf{x}'_{it}\boldsymbol{\beta}+u_i)}\right]^{y_{it}}\left[\frac{1}{1+\exp(\mathbf{x}'_{it}\boldsymbol{\beta}+u_i)}\right]^{1-y_{it}}f(u_i;\sigma^2)du_i\right)\text{。}$$

通过数值法可以求解似然函数，并将其视为 $\boldsymbol{\beta}$ 和 $\boldsymbol{\Sigma}$ 的函数，对其求最大值。处理这一问题的方法目前已有很多，接下来，我们将讨论最常见的几种方法。

12.6.2 Gauss-Hermite 求积分法

似然函数中求积分的维度取决于随机效应的结构。当维度较小时，如上述 logistic-正态模型（式 12.5）的单维积分，可以通过标准的数值积分法来对似然函数进行近似。

Gauss-Hermite 求积分法（Gauss-Hermite quadrature）是通过乘以另一个具有正态密度形状的函数，来对函数 $f(\cdot)$ 的积分求近似的方法。这个近似过程等于对在某些点上函数的解进行有限加权求和。在单维正态随机效应的情况下，该近似可以表示为

$$\int_{-\infty}^{\infty} f(u)\exp(-u^2)\,du \approx \sum_{k=1}^{q} c_k f(s_k),$$

其中*权数*为 $\{c_k\}$，*求积分点*（*quadrature point*）为 $\{s_k\}$。随着求积分点的数量 q 的增加，近似的效果也就越好。

通过标准算法如 Newton-Raphson 法可以对近似的似然函数求最大化，从而得出最大似然估计值 $\hat{\boldsymbol{\beta}}$ 以及 $\hat{\boldsymbol{\Sigma}}$。对所观察到的信息矩阵的近似值求逆，可以得出相应最大似然估计的标准误。对于复杂的模型，可以通过数值法而不是解析法来计算海塞（Hessian）矩阵的二阶偏导数。与 $\hat{\boldsymbol{\beta}}$ 相比，对标准误的充分近似往往要求更大的 q 值。在拟合时，我们建议逐步增大 q，直到估计值和标准误的变动都可以忽略不计为止。

一种可加形式的 Gauss-Hermite 求积分法（如：Liu and Pierce, 1994）将求积分点按照所要积分的函数的众数进行中心化处理（center），并根据在众数处所估计的曲率对其赋予相应的比例。这样做可以提高效率，大大降低对积分进行有效近似所需要的求积分点的数量。Lesaffre 和 Spiessens（2001）展示了相应的对比结果，并强调求积分点的数量不能太少。

12.6.3 蒙特卡洛法

多元形式的 Gauss-Hermite 求积分法可以处理多元和存在相关关系的随机效应。然而，当积分的维度大约超过 5 时，对似然函数进行充分的近似变得非常困难。这时，蒙特卡洛法在运算上比数值积分法更具有可行性。既有研究已经讨论过多种蒙特卡洛方法（如：McCulloch, 1997），其中包括与 Newton-Raphson 相结合的蒙特卡洛法、与 EM 算法相结合的蒙特卡洛法以及对似然函数进行直接模拟的蒙特卡洛法。在这里，我们简要介绍一种蒙特卡洛 EM（MCEM）算法。

当存在缺失数据或者补足某些"缺失"数据可以对似然函数进行简化时，EM 算法是一种常见的求解最大似然估计的迭代方法（Dempster et al., 1977）（有关综述，参见：Laird（1998））。每次迭代包括两个步骤：E 步（E-step）通过对缺失值求期望来近似似然函数，M 步（M-step）在给定参数估计的基础值后最大化似然函数。在广义线性混合模型中，可以将随机效应 $\mathbf{u}$ 视为缺失数据，那么，$h(\mathbf{y},\mathbf{u};\boldsymbol{\beta},\boldsymbol{\Sigma}) = f(\mathbf{y}\mid\mathbf{u};\boldsymbol{\beta})f(\mathbf{u};\boldsymbol{\Sigma})$ 指定了完整数据的联合分布。在 EM 算法中，第 r 次迭代的 E 步计算

$$E\left\{\log h(\mathbf{y},\mathbf{u};\boldsymbol{\beta},\boldsymbol{\Sigma}) \mid \mathbf{y};\boldsymbol{\beta}^{(r)},\boldsymbol{\Sigma}^{(r)}\right\}。$$

期望值是针对（$\mathbf{u}\mid\mathbf{y}$）的分布而言的，其中参数值等于第 r 次迭代的基础值 $\boldsymbol{\beta}^{(r)}$ 和 $\boldsymbol{\Sigma}^{(r)}$。根据贝叶斯定理，（$\mathbf{u}\mid\mathbf{y}$）的分布取决于广义线性混合模型中（$\mathbf{y}\mid\mathbf{u}$）和 $\mathbf{u}$ 的分布。接下来，M

步再将似然函数作为关于参数 $\boldsymbol{\beta}$ 和 $\boldsymbol{\Sigma}$ 的函数求最大化,以得出 $\boldsymbol{\beta}^{(r+1)}$ 和 $\boldsymbol{\Sigma}^{(r+1)}$。

MCEM 算法利用蒙特卡洛法对 E 步的期望进行近似。相应的方法包括:在当前的参数估计值处,根据 $(\mathbf{u}|\mathbf{y})$ 的分布进行独立的模拟;或者使用马尔科夫链蒙特卡洛(Markov chain Monte Carlo, MCMC)法。有关细节,包括如何确定一个蒙特卡洛样本的适当规模等问题,参见:Booth 和 Hobert(1999)、Chan 和 Kuk(1997)、McCulloch(1994,1997)。

12.6.4 惩罚性类似然近似

当更精致地使用 Gauss-Hermite 和蒙特卡洛积分法时(即,在数值积分法中增加求积分点的数量,在 MCEM 法中增大蒙特卡洛样本的规模),它们对似然函数的近似使所得到的参数估计收敛于最大似然估计。这一点与其他的近似方法不同,其他近似方法在计算上可能更为简便,但其估计结果却不一定接近于最大似然的估计结果。这些方法是对似然函数的一种解析近似结果进行最大化。

回顾式 12.18 似然函数,它是在 $\mathbf{y}$ 和 $\mathbf{u}$ 的联合分布中通过求积分剔除随机效应 $\mathbf{u}$ 所得的。利用该联合分布中每一项的指数族表达式,式 12.18 的被积函数可以表示为关于 $\mathbf{u}$ 的幂函数。对该函数进行近似的一种方法是对它的幂在一阶项等于 0 时对应的点 $\tilde{\mathbf{u}}$ 附近进行二阶泰勒级数展开(该点 $\tilde{\mathbf{u}} \approx E(\mathbf{u}|\mathbf{y})$)。这时,对被积函数的近似是关于 $(\mathbf{u}-\tilde{\mathbf{u}})$ 的二次项的幂函数,并且它具有对多元正态密度函数乘以一个常数的形式。因此,该函数的积分存在封闭解。这样的积分近似被称为*拉普拉斯近似*(*Laplace approximation*)。进而,将关于积分(式 12.18)的近似作为似然函数,并针对 $\boldsymbol{\beta}$ 和 $\boldsymbol{\Sigma}$ 求其最大值。

在拉普拉斯近似中,有一种积分近似(Breslow and Clayton, 1993)所得到的关于对数似然函数的近似函数为

$$q(\boldsymbol{\beta},\mathbf{y}) - (1/2)\tilde{\mathbf{u}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{u}},$$

其中 $q(\boldsymbol{\beta},\mathbf{y})$ 相当于以 $\mathbf{u}=\tilde{\mathbf{u}}$ 为条件的广义线性模型的类对数似然函数。因此,这个近似函数等于对类对数似然函数进行了一个惩罚,$\tilde{\mathbf{u}}$ 中的元素的绝对值越大,它的惩罚也就越大。这种方法被称为*惩罚性类似然法*(*penalized quasi-likelihood*, *PQL*)。通过有关针对结果变量服从正态分布的线性混合模型的方法,可以最大化惩罚性类似然函数。该方法将 Logit 的一个线性表达式作为操作性结果变量,并通过迭代法对一组关于 $\boldsymbol{\beta}$ 和 $\mathbf{u}$ 的类似然方程组求解。惩罚性类似然法不需要计算数值积分或蒙特卡洛积分,因而比最大似然法更为简便。对于大型数据以及随机效应结构非常复杂的模型来说,惩罚性类似然法在运算量上是可行的。

遗憾的是,与最大似然法相比,惩罚性类似然法的结果可能很差(McCulloch, 1997)。例如,对于第 12.3.2 节中有关流产态度的数据,惩罚性类似然法对最大似然估计的近似(由 SAS 的 GLIMMIX 得出)就 $\{\beta_t\}$ 来说还算不错,但是它对标准误以及 σ 的估计仅为最大似然估计的一半左右(比如,惩罚性类似然法得出 $\hat{\sigma}=4.3$,而相应的最大似然估计为 8.6)。当真实的方差构成很大时,普通的惩罚性类似然法一般会出现严重低估(Breslow and Lin, 1995)。另外,当结果变量的分布严重偏离正态分布时(如二分变量),惩罚性类似然估计结果也会存在很大问题。目前,已经发展出了一些调整方法以降低有关的偏差(如:Goldstein and Rasbash, 1996)。但是我们建议,如果有可能的话,应尽量使用最大似然法。

12.6.5 贝叶斯方法

拟合广义线性混合模型的另一种方法是贝叶斯方法。对这种方法来说,关于固定效

应和随机效应的区分不再存在,因为每个效应都具有一个概率分布。利用一个扁平的先验分布得出后验分布,它等于似然函数乘上一个常数。然后,通过马尔科夫链蒙特卡洛(Markov chain Monte Carlo,MCMC)法对难以处理的后验分布进行近似,便可以得到关于似然函数的近似(Zeger and Karim, 1991)。例如,对后验分布的众数的近似就是对有关最大似然估计的近似。

使用该方法的风险在于,对许多关于分类数据的模型来说,不适当的先验分布会导致不恰当的后验分布(Natarajan and McCulloch, 1995)。在应用 MCMC 法时,我们可能无法认识到后验分布是不恰当的。因而,更安全的做法是,选用一个适当的但相对分散的先验分布。但是,后验分布的众数并不一定就与最大似然估计相近,并且马尔科夫链可能会收敛得很慢(Natarajan and McCulloch, 1998)。目前,这仍然是一个热点的研究领域,它不仅仅是对最大似然结果进行近似的一种方法,甚至贝叶斯学派的统计学家认为这种方法要优于最大似然法。比如,参见 Daniels 和 Gatsonis(1999)关于群组跟踪二分数据的时空趋势的多层次模型分析,该研究参考了 Wong 和 Mason(1985)有关层级模型的早期成果。

12.6.6 模型参数的推断

在完成模型拟合后,有关固定效应的推断方法与普通模型相同。例如,可以利用似然比检验来比较嵌套模型。关于广义线性混合模型的渐近推断随着群组数量而不是组内观测值数量的增加而变得有意义。相似地,在使用以大量群组为基础的再抽样方法如重抽样自举法(bootstrap)时,应当抽取群组单位而不是组内的个体观测值,以保持组内的相依性。

关于随机效应(如它们的方差结构)的推断相对更复杂一些。例如,有时一个模型是另一个模型在方差构成等于 0 时的特例。这种情况下,较简单的模型会落在相对复杂的模型的参数空间的边界上,因而无法应用基于似然的普通推断方法。在大多数情况下,用来检验一个只包括单一方差构成项的模型中 $H_0:\sigma^2=0$(备择假设为 $H_a:\sigma^2>0$)的似然比统计量的渐近分布是已知的。它的零分布是关于 χ_0^2(即在 0 处退化)和 χ_1^2 随机变量的均匀混合(Self and Liang, 1987)。零假设 $\sigma^2=0$ 的情况发生在当 $\hat{\sigma}=0$ 时,这时在 H_0 和 H_a 下的最大化似然函数完全相同。当 $\hat{\sigma}>0$ 且所观察到的检验统计量等于 t 时,大样本检验的 P 值为 $\frac{1}{2}P(\chi_1^2>t)$,即等于使用 χ_1^2 渐近检验时对应的 P 值的一半。当对多个方差构成项进行检验时,相应的混合分布变得更加复杂,因而使用计分检验相对更为简便(Lin, 1997)。

12.6.7 通过随机效应进行预测

在模型中使用随机效应意味着,诸如发生比之比等研究所关注的效应具有异质性。这时,所估计的效应通常是有关固定效应和随机效应的线性组合。例如,在比较两种治疗方式的临床试验中包括中心别随机效应的模型(第 12.3.4 节),我们可以预测在每个中心每种治疗方式的成功概率以及各中心之间的发生比之比。

给定这些数据,$(\mathbf{u}|\mathbf{y})$ 的条件分布包含了有关随机效应 $\mathbf{u}$ 的信息。对 $\mathbf{u}$ 的一个预测为 $E(\mathbf{u}|\mathbf{y})$,即给定数据情况下它的*后验均值*(*posterior mean*)。对 $E(\mathbf{u}|\mathbf{y})$ 的计算需要使用数值积分或蒙特卡洛近似。该期望值取决于 $\boldsymbol{\beta}$ 和 $\boldsymbol{\Sigma}$,因而,在应用中可以将 $\hat{\boldsymbol{\beta}}$ 和 $\hat{\boldsymbol{\Sigma}}$ 代入近似过程。关于随机效应 u_i 的预测值的标准误等于 $(u_i|\mathbf{y})$ 的分布的标准差。然而,当我们将 $\hat{\boldsymbol{\beta}}$ 和 $\hat{\boldsymbol{\Sigma}}$ 代入 $E(\mathbf{u}|\mathbf{y})$ 时,标准误并没有包括这些估计值本身的样本变动性。因此,

我们一般会低估真实的标准误(Booth and Hobert, 1998)。

这种利用随机效应的后验均值进行预测的方法所提供的效应估计,是对仅使用某个群组内的数据进行估计的一种收缩。从这个角度来说,该结果与利用*经验贝叶斯*(*empirical Bayes*)方法所得到的结果相似(Ten Have and Localio, 1999)。经验贝叶斯方法是一种利用样本数据来估计先验分布中的参数的普通贝叶斯分析。对于一个由参数均值组成的向量,这种方法所得到的关于某个特定均值的估计等于对该样本均值以及基于所有样本均值的总体均值的一种加权平均。因此,它将样本均值向总体均值收缩。当用来估计每个参数的样本规模很小、需要估计许多参数或者参数的真值大致相等时,收缩估计值(shrinkage estimator)大大优于样本值。经验贝叶斯理论已经在一些场合得到了广泛应用:例如,用于估计一个关于均值或二项分布比例的向量(Efron and Morris, 1975)。

尽管在许多现实应用中,随机效应模型是一种自然的选择,但是有关的进一步研究工作十分必要。关于复杂广义线性混合模型的模型拟合以及统计推断的方法仍在进一步探讨之中。另外,有关模型检查和诊断的技术也有待发展。但是,我们坚信,广义线性混合模型是对普通广义线性模型的一个非常重要的扩展。

注解

第 12.1 节:关于群组分类变量的随机效应模型

12.1 有关 Rasch 模型及其参数拟合方法的详细讨论,参见:Anderson(1980, Sec. 6.4)、Fischer 和 Molenaar(1995)。Haberman(1977b)表明,当 n 和 T 都按照适当的比率增加时,最大似然估计值满足一致性。有关 Rasch 模型的多项分布扩展,参见:Anderson(1980, pp. 272-284;1995)、Conaway(1989)。关于分类结果变量的随机效应模型的早期研究包括:Anderson 和 Aitkin(1985)、Bartholomew(1980)、Bock 和 Aitkin(1981)、Chamberlain(1980)、Gilmour 等(1985)、Pierce 和 Sands(1975)、Stiratelli 等(1984)。

12.2 Neuhaus 和 Lesperance(1996)指出,在包括协变量的模型中,当群组规模较小并且协变量存在很强的组内正相关时,条件最大似然法与随机效应方法相比可能会损失相当大的效率。随着相关系数趋近于 +1,协变量的效应类似于一个组间效应,而这是条件最大似然法所无法估计的。在第 12.1.2 节所提到的配对数据中,条件最大似然估计与随机效应估计相等,其中组内协变量相关系数等于 -1,基于观测值之间的不同排序,x_t 的值从 0 变为 1 或从 1 变为 0,因而,这里不存在效率损失。

第 12.3 节:二分数据随机效应模型的例子

12.3 有关对捕获-再捕获数据进行模型分析的详细讨论,参见:Bishop 等(1975, Chap. 6)、Chao 等(2001)、Cormack(1989)、Coull 和 Agresti(1999)、Darroch 等(1993)、Fienberg 等(1999)、Hook 和 Regal(1995)。这个问题与在观察到独立的 bin(n,π)计数情况下估计二项分布基数 n 的相应问题很相似,其中 n 和 π 未知,参见:Aitkin 和 Stasinopoulos(1989),以及文中的参考文献。允许捕获概率存在异质性的其他模型,也可能出现对数似然函数相对扁平的情况(Burnham and Overton, 1978),比如 β-二项模型。

12.4 对于在分析集合的分类数据时所遇到的问题——*生态推断*(*ecological inference*),King(1997)利用随机效应模型作为一种解决方法。Chambers 和 Steel(2001)对

Leo Goodman 在这方面的早期工作进行了讨论，并提出了一个更简单的半参数方法。

第 12.4 节：多项分布数据的随机效应模型

12.5 在补余双对数连结的情况下，似然函数具有封闭形式的解，其中随机效应服从对数 γ 分布(Crouchley, 1995, Farewell, 1982, Ten Have, 1996)。

12.6 Chen 和 Kuo(2001)讨论了定类结果变量的情况，包括带有随机效应的离散选择模型(第 7.6 节)。有关离散广义线性混合模型的讨论，另见：Brownstone 和 Train (1999)。

第 12.5 节：二分数据的多元随机效应模型

12.7 Rabe-Hesketh 和 Skrondal(2001)指出，在包括多元随机效应的模型中，应特别注意参数的可识别性。他们的因子分析模型包含了大量多元随机效应模型作为特例。

12.8 对于跟踪数据中的二元二分结果变量，Ten Have 和 Morabia(1999)对两个变量的对数发生比之比和单维 Logit 进行了同步模型分析。有时，多元结果变量既包括连续变量，也包括分类变量。有关分析这类数据的随机效应模型，参见：Catalano 和 Ryan(1992)、Gueorguieva 和 Agresti(2001)。

第 12.6 节：广义线性混合模型的拟合、推断与预测

12.9 有关拟合广义线性混合模型的更多细节，参见：Fahrmeir 和 Tutz(2001, Chap. 7)、McCulloch 和 Searle(2001)。正如广义线性混合模型的似然函数具有积分形式，它的似然方程也具有积分方程的形式(McCulloch and Searle, 2001, p. 227)。Wolfinger 和 O'Connell(1993)描述了一种与惩罚性类似然法有关的拟合方法，该方法同样使用了拉普拉斯近似。

12.10 广义线性混合模型决定了结果变量的均值与解释变量之间的边际关系(对随机效应求平均)。与之相对，Heagerty(1999)指出，关于均值的边际模型隐含决定了条件模型中线性预测项的固定效应部分。条件广义线性混合模型(式 12.1)的线性预测项为 $\mathbf{x}_{it}'\boldsymbol{\beta}+\mathbf{z}_{it}'\mathbf{u}_i$，其更一般的表述形式 $\Delta_{it}+\mathbf{z}_{it}'\mathbf{u}_i$ 意味着某个特定的边际模型。在这里，Δ_{it} 是边际模型的线性预测项以及随机效应分布的一个函数。它由连结边际均值和条件均值的积分方程隐含决定。

习题

应用部分

12.1 参见表 10.14 和习题 10.1 的配对数据。

a. 拟合式 12.3 模型，并解释 $\hat{\beta}$。如果你的软件使用的是数值积分法，报告当求积分点分别为 5，10，25，100 和 200 时对应的 $\hat{\beta}$，$\hat{\sigma}$ 以及它们的标准误，并对收敛的情况加以评述。

b. 将使用这种方法得到的 $\hat{\beta}$ 及其标准误与条件最大似然法的相应估计结果进行比较。

12.2 参见表 4.8 中有关沙克·奥尼尔的罚球数据。在第 i 场比赛中，假定 y_i = 在总共 n_i 次罚球中罚中的次数，它是一个 $\text{bin}(n_i, \pi_i)$ 变量，并且 $\{y_i\}$ 之间相互独立。

a. 拟合模型 $\text{Logit}(\pi_i)=\alpha$。求 $\hat{\pi}_i$ 并对其加以解释。这个模型对数据拟合得充分吗？

b. 拟合模型 $\text{Logit}(\pi_i)=\alpha+u_i$，其中 $\{u_i\}$ 服从独立的 $N(0,\sigma^2)$。利用 $\hat{\alpha}$ 和 $\hat{\sigma}$ 描述

奥尼尔的罚球情况。

c. 说明为什么(a)部分的模型是(b)部分模型的一个特例。有没有证据表明(b)部分的模型拟合得更好?

12.3 对于表8.3的数据,当对象 i 有过行为 t 时,令 $y_{it}=1$。表12.11给出了logistic-正态模型

$$\text{Logit}\left[P(Y_{it}=1 \mid u_i)\right]=u_i+\beta_t$$

对该数据的拟合结果。通过比较吸烟和服用大麻的情况来加以说明。

表12.11 习题12.3的输出结果

Description	Value	Parameter	Estimate	Std Error	t Value
Subjects	2 276				
Max Obs Per Subject	3	beta1	4.222 7	0.182 4	23.15
Parameters	4	bata2	1.620 9	0.120 7	13.43
Quadrature Points	200	bata3	−0.775 1	0.106 1	−7.31
Log Likelihood	−3 311	sigma	3.549 6	0.162 7	21.82

12.4 习题12.3中的模型与第8.2.4节所使用的对数线性模型(AC,AM,CM)的关注点有何不同?如果 $\hat{\sigma}=0$,哪一个对数线性模型具有与这个广义线性混合模型相同的拟合结果。

12.5 对于表9.1的学生调查数据,(a)利用广义线性混合模型分析该数据,(b)将结果与习题11.2中边际模型的有关结果加以比较并进行解释。

12.6 对流产态度的数据拟合式12.10模型。如果你的软件使用的是Gauss-Hermite求积分法,报告参数拟合收敛以及标准误拟合收敛所需要的求积分点的大致数量(这个例子中的 $\hat{\sigma}$ 很大,因而需要许多求积分点)。

12.7 对于表11.10中的交叉研究,拟合模型

$$\text{Logit}\left[P(Y_{i(k)t}=1 \mid u_{i(k)})\right]=\alpha_k+\beta_t+u_{i(k)}, \tag{12.19}$$

其中$\{u_{i(k)}\}$服从独立的$N(0,\sigma^2)$。对$\{\hat{\beta}_t\}$和 $\hat{\sigma}$ 加以解释。

12.8 在习题12.7中,比较下列模型所估计的 $\beta_B-\beta_A$,$\beta_C-\beta_A$ 及其标准误的值:(a)边际模型(习题11.6);(b)条件logistic回归(第10.2节),将式12.19模型中对象层面的项视为固定效应。

12.9 对于习题12.7,拟合一个允许干预效应$\{\beta_{tk}\}$随着干预次序的变动而变动的更一般的广义线性混合模型。检验这个模型是否拟合得更好。我们也可以考虑时期效应或延滞效应(carryover effects)。在式12.19模型中,加入两个时期效应(比如,当 $t=A$ 且 $k=1,2$,$t=B$ 且 $k=3,4$,以及 $t=C$ 且 $k=5,6$ 时,将第一个时期效应参数加入模型)。检验这个模型是否拟合得更好。对结果加以解释。

12.10 对于有关流产态度的数据,考虑logistic-正态模型(式12.10),限定条件为 $\sigma=0$。

a. 说明为什么拟合结果与将每个调查对象的三次回答视为三个不同对象的独立回答的普通Logit模型结果相同。

b. 说明为什么模型拟合结果与关于在给定 G=性别后三个回答结果(I_1,I_2,I_3)相互独立的普通对数线性模型(GI_1,GI_2,GI_3)相同。

c. 拟合该模型。解释结果,并说明为什么$\{\hat{\beta}_t - \hat{\beta}_u\}$与第12.3.2节中允许$\sigma > 0$的情况存在很大差别。

12.11 关于表6.7中研究生录取决定的数据,令$y_{ig} = 1$表示一个申请第i个系的性别为g(1 = 女性,0 = 男性)的对象被录取了。

a. 在固定效应模型 $\text{Logit}[P(Y_{ig} = 1)] = \alpha + \beta g + \beta_i^D$ 中,$\hat{\beta} = 0.173$(SE = 0.112)。加以解释。

b. 相应的在系别层面上具有正态随机效应的式12.12模型中,$\hat{\beta} = 0.163$(SE = 0.111)。加以解释。

c. 在式12.12模型中允许不同系别具有不同的性别效应,结果$\hat{\beta} = 0.176$(SE = 0.132),$\hat{\sigma}_b = 0.20$。加以解释。说明为什么$\hat{\beta}$的标准误比其他的分析结果更大了一些。

d. 在边际分布中,性别与是否被录取之间的样本对数发生比之比等于-0.07。为什么它的符号与这些模型中的$\hat{\beta}$相反?

e. 性别与是否被录取之间的样本条件发生比之比的变动范围为0到∞。与之相比,在包括交互项的随机效应模型中,所预测的发生比之比的变动范围很小。说明为什么会出现这么大的差异。

12.12 对于表9.16中的临床试验数据,令$\pi_{it} = P(Y_{it} = 1 | u_i)$表示在第$i$个中心干预方式为$t$的成功概率。

a. 随机截距模型(式12.11)的拟合结果为$\hat{\beta} = 1.52$(SE = 0.70)以及$\hat{\sigma} = 1.9$。加以解释。

b. 根据第9.8.3节,这个模型所对应的固定效应模型(将$\alpha + u_i$替代为α_i)具有$\hat{\sigma}_1 = \hat{\sigma}_3 = -\infty$,即对每种干预方式$\hat{\pi}_{1t} = \hat{\pi}_{3t} = 0$。相反,随机效应模型具有$\hat{\sigma} + \hat{u}_1 = -3.78$(由SAS的NLMIXED程序给出),第1个中心的$\hat{\pi}_{11} = 0.047$,$\hat{\pi}_{12} = 0.011$。说明为什么在这个模型中,对于没有成功案例的中心会出现$\hat{\pi}_{it} > 0$。

12.13 参考第12.3.3节的对象别模型。验证对于治疗抑郁症的新药和标准药之间所估计的时间效应的斜率之差:(a)在广义线性混合模型方法中等于1.018(SE = 0.192);(b)在条件最大似然法中等于1.156(SE = 0.222)。

12.14 对表10.5的婚前和婚外性行为数据,与所拟合的边际模型(式10.14)相对应,表12.12给出了随机截距模型的拟合结果。解释$\hat{\beta}$。将这里关于β的估计和推断结果与第10.3.2节中边际模型的相应结果加以比较。

表12.12 习题12.14的输出结果

				Std	
Subjects	475	Parameter	Estimate	Error	t Value
Max Obs Per Subject	2	inter1	-1.542 2	0.182 6	-8.45
Parameters	5	inter2	-0.668 2	0.157 8	-4.24
Quadrature Points	100	inter3	0.927 3	0.167 3	5.54
Log Likelihood	-890.1	beta	4.134 2	0.329 6	12.54
		sigma	2.075 7	0.248 7	8.35

12.15　1994 年美国综合社会调查(General Social Survey)收集了调查对象对政府是否应该增加、保持不变或减少四项(环境、健康、执法、教育)投入的态度。数据中还包括每个对象的性别和种族。对于第 i 个对象,令 $G_i=1$ 表示女性、0 表示男性,$R_{1i}=1$ 表示白人、0 表示非白人,$R_{2i}=1$ 表示黑人、0 表示非黑人,$R_{1i}=R_{2i}=0$ 表示其他种族。令 y_{it} 表示第 i 个对象关于第 t 项政府投入的态度,这里结果(1,2,3)分别代表(增加,保持不变,减少)。

a. 在限定条件为 $\beta_4=0$ 的情况下,随机截距模型

$$\text{Logit}\left[P(Y_{it}\leqslant J\mid u_i)\right]$$
$$=\alpha_j+\beta_t+\beta_g G_i+\beta_{r1}R_{i1}+\beta_{r2}R_{2i}+u_i, j=1,2$$

具有 $\hat{\beta}_1=-0.55$, $\hat{\beta}_2=-0.60$, $\hat{\beta}_3=-0.49$,且 $\hat{\sigma}=1.03$。这些估计值的绝对值均大于其标准误的五倍。对结果加以解释。

b. 表 12.13 给出了种族与政府投入项目之间交互项的拟合结果。加以解释。

表 12.13　习题 12.15 的结果[a]

变　量	参数估计	SE(标准误)
截距-1	1.065	0.391
截距-2	1.919	0.051
性别	0.409	0.088
种族 1-白人	-0.055	0.397
种族 2-黑人	0.434	0.452
项目 1-环境	-0.357	0.539
项目 2-健康	-0.319	0.493
项目 3-犯罪	-0.585	0.480
种族 1 * 项目 1	-0.170	0.549
种族 1 * 项目 2	-0.387	0.503
种族 1 * 项目 3	0.197	0.491
种族 2 * 项目 1	-0.452	0.606
种族 2 * 项目 2	0.454	0.598
种族 2 * 项目 3	-0.518	0.560

a 第 4 项政府投入(教育)和第 3 类种族(其他)的编码为 0。

12.16　参考习题 11.12 中表 8.19 关于政府花费的数据。利用一个包括随机效应的累积 Logit 模型来分析这些数据。对结果加以解释,并与边际模型的相应结果(习题 11.12)进行对比。

12.17　对于第 12.4.2 节中有关失眠症的例子,SAS 给出的最大化对数似然值等于 -593.0,令 $\sigma=0$ 的较简单模型的相应值为 -621.0。通过似然比检验或者 AIC 来比较这些模型。你的结论是什么?

12.18　Landis 和 Koch(1977)给出了七位病理学家分别对 118 张有关子宫颈癌严重程度的图片进行评定的结果,该结果由一个五分的刻度来表示(表 13.1 是将他们的表格的前两类和后三类分别进行合并后的简化版)。对于第 i 张图片和第 t 个评定者,表 12.14 给出了对定序表格拟合模型

$$\text{Logit}\left[P(Y_{it}\leqslant j\mid u_i)\right]=u_i+\alpha_j+\beta_t$$

的结果(令 $\hat{\beta}_G=0$),其中假定 $\{u_i\}$ 服从独立的 $N(0,\sigma^2)$。表中还给出了相应的边

际模型的 GEE 估计结果，使用独立性的相关结构设定。解释每个模型中的 $\hat{\beta}_F$。说明为什么随机效应模型的估计值的绝对值要大得多，这里 $\hat{\sigma}=3.8$。讨论这两个模型在模型假设以及对结果的解释方面的差别。

表 12.14 习题 12.18 的结果

评定者	GEE	随机效应
A	-0.451(0.108)	-1.201(0.300)
B	-0.391(0.093)	-0.919(0.299)
C	0.319(0.118)	0.558(0.301)
D	0.632(0.105)	1.545(0.313)
E	-0.491(0.098)	-1.379(0.300)
F	1.252(0.161)	2.907(0.344)

12.19 参考在第 12.5.1 节中学校男生对前卫群体认同态度的数据。表 12.15 给出了一个女生样本的结果。拟合式 12.16 模型，并加以解释。综述所估计的随机效应的变动性及其相关结构。

表 12.15 习题 12.19 的数据

第一次调查时的(*M*,*A*)	第二次调查时的(*M*,*A*)[a]			
	(属于，正面)	(属于，负面)	(不属于，正面)	(不属于，负面)
属于，正面	484	93	107	32
属于，负面	112	110	30	46
不属于，正面	129	40	768	321
不属于，负面	74	75	303	536

a *M*：成员认同；*A*：态度。

来源：J. S. Coleman, *Introduction to Mathematical Sociology* (London: Free Press of Glencoe, 1964), p. 168。

12.20 对式 12.16 模型进行扩展，以同时分析表 12.8 和表 12.15 的数据，在模型中包括一个关于性别的主效应，并令每个性别都具有相同的成员认同效应和态度效应。拟合该模型。通过最大化对数似然值将其与一个允许不同性别具有不同的成员认同效应和不同态度效应的更一般化模型进行对比。对结果加以解释。

12.21 表 12.16 取自一项对 1995 年中国台湾甲型肝炎流行中被感染人数 *N* 进行估计的研究。研究共报告了 271 个病例，分别来自于中国台湾预防疾病研究所(P)的血清检测、检疫服务中心(Q)的记录，以及流行病学家的问卷调查(E)。这里估计 *N* 非常困难，因为许多病例只“被捕获”了一次。

表 12.16 习题 12.21 的数据

PEQ	观察计数	Logistic-正态模型的最大似然拟合值
000	—	(487, ∞)
001	63	61.0
010	55	58.0
011	18	17.0
100	69	68.0
101	17	20.0
110	21	19.0
111	28	28.0

来源：数据取自：Chao et al. (2001)。

a. 在以下条件下分别求 $\hat{N}$:(i)只观察到 P 和 Q,(ii)只观察到 P 和 E,(iii)只观察到 Q 和 E。

b. 通过关于 P,Q 和 E 之间的相互独立性模型求 $\hat{N}$。

c. 利用(b)部分的模型求关于 N 的 95% 的剖面似然区间。

d. 表 12.16 给出了第 12.3.6 节所介绍的随机效应模型的拟合结果,其中 $\hat{\sigma}$ = 2.9。对数似然函数相对扁平,$\hat{N}$ = 4 551 的 95% 的剖面似然区间为(758,∞)(Coull and Agresti, 1999)。说明为什么这个模型给出的关于 N 的估计可能不精确。由于(c)部分所得的区间要小得多,那么是不是它就一定更可靠呢?

12.22 利用随机效应方法分析表 11.1 中交叉研究的数据。解释结果,并将其与第 11.1.2节的相应结果进行对比。

12.23 第 12.3.2 节中有关态度比较的分析可以扩展到定序结果变量的情况。利用定序随机效应模型,分析 Agresti(1993)中的 4^3 表格,该表格也可从本书的网站(*www.stat.ufl.edu/~aa/cda/cda.html*)上找到。

12.24 第 12.3.4 节有关多中心临床试验中异质性的分析可以扩展到定序结果变量的情况。通过随机效应模型,分析 Hartzel 等(2001a)中的 2×3×8 表格。

12.25 假如你是一个统计学顾问,被要求分析 B. Efron(*Statistical Science* 13:95-122, 1998)中的表 4,它来自于一项在 41 个城市进行的临床试验的 2×2 表格。分析这些数据,并完成一篇数据分析报告。

12.26 以年龄和母亲吸烟状况作为预测变量,利用下列模型分析表 11.9:(a)logistic-正态模型,(b)边际模型,(c)转换模型。说明为什么在这些模型中对母亲吸烟效应的解释有所不同。

理论与方法

12.27 参考第 12.3.1 节。利用辅助信息可以提高预测能力。令 q_i 表示在 1992 年大选中,以投票给克林顿或布什为条件,投票给克林顿的真实比例。考虑模型

$$\text{Logit}\left[P(Y_{it}=1\mid u_i)\right]=\text{Logit}(q_i)+\alpha+u_i,$$

其中 $\{q_i\}$ 是已知的,并且 $\{u_i\}$ 服从独立的 $N(0,\sigma^2)$。当 $\hat{\sigma}=0$ 时,证明 $\hat{\pi}_i=q_i\exp(\hat{\alpha})/[1-q_i+q_i\exp(\hat{\alpha})]$。将其与 $\{q_i\}$ 进行比较,说明 $\hat{\pi}_i$ 如何随着民主党在当前民意调查中与以往选举相比的总得票情况而上下波动(即取决于 $\hat{\alpha}$)。证明当 $\hat{\alpha}=0$ 时,$\hat{\pi}_i=q_i$。

12.28 对于一个二分结果变量,考虑随机效应模型

$$\text{Logit}\left[P(Y_{it}=1\mid u_i)\right]=\alpha+\beta_t+u_i,\quad t=1,\cdots,T,$$

其中 $\{u_i\}$ 服从独立的 $N(0,\sigma^2)$,以及边际模型

$$\text{Logit}\left[P(Y_t=1)\right]=\alpha+\beta_t^*,\quad t=1,\cdots,T。$$

考虑到模型的可识别性,令 $\beta_T=\beta_T^*=0$。说明为什么所有 $\beta_t=0$ 意味着所有 $\beta_t^*=0$。反过来成立吗?

12.29 对二分数据使用 probit 连结函数的广义线性混合模型为

$$\Phi^{-1}\left[P(Y_{it}=1\mid \mathbf{u}_i)\right]=\mathbf{x}_{it}'\boldsymbol{\beta}+\mathbf{z}_{it}'\mathbf{u}_i,$$

其中 Φ 是 $N(0,1)$ 的累积分布函数，并且 $\mathbf{u}_i$ 满足 $N(\mathbf{0},\mathbf{\Sigma})$ 的概率密度函数 $f(\mathbf{u}_i;\mathbf{\Sigma})$。

a. 证明边际均值等于

$$P(Y_t = 1) = \int P(Z - \mathbf{z}'_{it}\mathbf{u}_i \leqslant \mathbf{x}'_{it}\boldsymbol{\beta})f(u_i;\mathbf{\Sigma})\,d\mathbf{u}_i\ ,$$

其中 Z 是一个独立于 $\mathbf{u}_i$ 的标准正态变量。

b. 由于 $Z - \mathbf{z}'_{it}\mathbf{u}_i$ 服从 $N(0,1+\mathbf{z}'_{it}\mathbf{\Sigma}\mathbf{z}_{it})$ 分布，推出

$$\Phi^{-1}\left[P(Y_t = 1)\right] = \mathbf{x}'_{it}\boldsymbol{\beta}\left[1 + \mathbf{z}'_{it}\mathbf{\Sigma}\mathbf{z}_{it}\right]^{-1/2}。$$

因而，边际模型是一个效应被弱化后的 probit 模型。以单维随机截距模型为例，证明边际效应等于广义线性混合模型中的相应效应除以 $\sqrt{1+\sigma^2}$。

12.30 在 Rasch 模型 $\text{Logit}[P(Y_{it}=1)] = \alpha_i + \beta_t$ 中，α_i 是固定效应。

a. 假定不同对象的回答以及同一对象的不同观测值之间都相互独立，证明对数似然函数是

$$\sum_i\sum_t \alpha_i y_{it} + \sum_i\sum_t \beta_t y_{it} - \sum_i\sum_t \log\left[1 + \exp(\alpha_i + \beta_t)\right]。$$

b. 对于所有 i 和 t，证明似然方程为 $y_{+t} = \sum_i P(Y_{it} = 1)$ 和 $y_{i+} = \sum_t P(Y_{it} = 1)$。说明为什么以 $\{y_{i+}\}$ 为条件所得出的分布不取决于 $\{\alpha_i\}$。

c. 相应地，讨论将 α_i 视为随机效应的优缺点。

12.31 考虑关于配对数据的随机效应模型（式 12.3）。对于给定的 β_0，δ_0 是使得 $\hat{\mu}_{12} = n_{12} + \delta_0$ 和 $\hat{\mu}_{21} = n_{21} - \delta_0$ 满足 $\log(\hat{\mu}_{21}/\hat{\mu}_{12}) = \beta_0$ 的值。假定 $\{\hat{\mu}_{ij}\}$ 具有非负的对数发生比之比，说明为什么：

a. 这是假定 $\beta = \beta_0$ 时的模型拟合结果。

b. 这个模型中检验 $H_0: \beta = \beta_0$ 的似然比统计量为

$$2\left(n_{12}\log\frac{n_{12}}{n_{12} + \delta_0} + n_{21}\log\frac{n_{21}}{n_{21} - \delta_0}\right)。$$

c. 关于 $H_0: \beta = 0$ 的似然比检验就是相应的对称性检验。

12.32 说明为什么当在捕获-再捕获实验中只进行两次捕获时，logistic-正态模型没有实际意义。

12.33 参考习题 12.7 中的交叉研究。利用定序尺度（没有，中度，完全）表示症状缓解状况，Kenward 和 Jones（1991）给出了相应的分析结果。说明如何对这些数据构建一个类似于式 12.19 模型的定序 Logit 随机效应模型。

12.34 利用相邻类别 Logit 构建一个与累积 Logit 模型（式 12.14）相同的模型。解释其中的参数。

12.35 对于计数为 $\{n_{ab}\}$ 的 $I \times I$ 的定序方形表格，有关对象 i 的二分配对结果变量（Y_{i1}，Y_{i2}）的模型（式 12.3）可以扩展为

$$\text{Logit}\left[P(Y_{it} \leqslant j \mid u_i)\right] = \alpha_j + \beta x_t + u_i\ ,$$

其中 $\{u_i\}$ 是独立的 $N(0,\sigma^2)$ 变量，并且 $x_1 = 0$，$x_2 = 1$。

a. 说明如何解释 β，并将其与相应的式 10.14 边际模型中的 β 进行对比。

b. 这个模型隐含着关于将类别 1 到 j 和类别 $j+1$ 到 I 分别合并后所得到的每个 2×2 表格构建的式 12.3 模型。利用二分配对数据的条件最大似然（或随机效

应最大似然)估计值,说明为什么

$$\log\left[\left(\sum_{a>j}\sum_{b<j}n_{ab}\right)/\left(\sum_{a<j}\sum_{b>j}n_{ab}\right)\right]$$

是对β的一致性估计。

c. 暂且将合并后的$(I-1)$个2×2表格视为独立的样本,证明将每个e^{β}的估计值的分子和分母分别加总,可得出关于β的概括估计值(summary estimator)

$$\tilde{\beta}=\log\left\{\left[\sum_{a>b}(a-b)n_{ab}\right]/\left[\sum_{b>a}(b-a)n_{ab}\right]\right\}。$$

说明为什么即使考虑到各表间实际的相依关系,$\tilde{\beta}$仍然是关于β的一致性估计。

d. 在计算关于$\tilde{\beta}$的标准误时,将(c)部分的2×2表格视为相互独立是不正确的。将$\{n_{ab}\}$视为一个多项分布样本,证明所估计的关于$\tilde{\beta}$的渐近方差为(Agresti and Liang, 1993a)

$$\left\{\sum_{b>a}(b-a)^2 n_{ab}/\left[\sum_{b>a}(b-a)n_{ab}\right]^2\right\}+$$
$$\left\{\sum_{a>b}(a-b)^2 n_{ab}/\left[\sum_{a>b}(a-b)n_{ab}\right]^2\right\}。$$

12.36 总结广义线性混合模型方法相对于边际模型方法的优缺点。分别描述在下列方法中参数估计满足一致性的条件:(a)使用GEE法的边际模型,(b)使用最大似然法的边际模型,(c)使用惩罚性类似然法(PQL)的广义线性混合模型,(d)使用最大似然法的广义线性混合模型。

13 关于分类数据的其他混合模型*

在第 10 至 12 章中，我们介绍了分析由于重复测量或其他形式的群组现象所导致的观测值之间相关的数据的方法。第 12 章所讨论的广义线性混合模型（generalized linear mixed models，GLMMs）假定随机效应服从正态分布，利用一系列服从正态分布的线性预测项的混合体来替代普通的线性预测项，从而描述了异质性。在本章中，我们介绍一些与广义线性混合模型有关的其他模型。除其中一种情况以外，本章所介绍的模型使用的均为非正态的混合分布。

第 13.1 节介绍潜类模型（latent class models）。这种模型将列联表视为由多个未观察到的表格混合而成的，这些未观察到的表格分别对应于一个定性潜变量的不同取值。在第 13.2 节中，我们讨论拟合广义线性混合模型的一种非参数方法，该方法利用一个未设定的离散定量分布来表示随机效应的分布。

在第 13.3 节，我们对群组的二项分布结果变量进行模型分析，并利用 β 分布（beta distribution）来描述二项分布参数之间的异质性。值得指出的是，由此得到的 β-二项分布（beta-binomial distribution）的方差函数也适用于类似然法。在第 13.4 节，我们对计数结果变量进行模型分析，并通过 γ 分布（gamma distribution）来描述泊松分布参数之间的异质性。由此得到的负二项（negative binomial）回归模型对应于一个随机效应服从对数-γ 分布（log-gamma distribution）的泊松广义线性混合模型。对服从泊松分布的结果变量来说，除了具有正态随机效应的广义线性混合模型之外，负二项回归模型也是一种可选的模型，对此我们将在第 13.5 节加以讨论。

13.1 潜类模型

广义线性混合模型通过一个服从正态分布的潜变量，即未观察到的随机效应向量，生成混合的线性预测项。与之相反，潜类模型使用一个定性的而不是定量的混合分布。其基本模型假定存在一个分类潜变量，在给定该变量取值后，所观察到的结果变量条件独立。

对于分类结果变量 $(Y_1, Y_2, \cdots, Y_T)$，潜类模型假定存在一个分类潜变量 Z，使得在 Z 的每个类别 z 处，结果变量的每个可能的取值组合 $(y_1, \cdots, y_T)$ 满足

$$P(Y_1 = y_1, \cdots, Y_T = y_T \mid Z = z) = P(Y_1 = y_1 \mid Z = z) \cdots P(Y_T = y_T \mid Z = z)。$$

图 13.1 显示了该模型的关联图。潜类模型分析的是每个潜类发生的概率 $P(Z = z)$ 以及在一个潜类内关于 Y_t 的条件概率 $P(Y_t = y_t \mid Z = z)$，这些都是模型的参数。更一般的情况下，潜变量 Z 可以是多元的。潜类模型与有关多元正态结果变量的因子分析模型相类

似,不过它针对的是分类结果变量和潜变量。

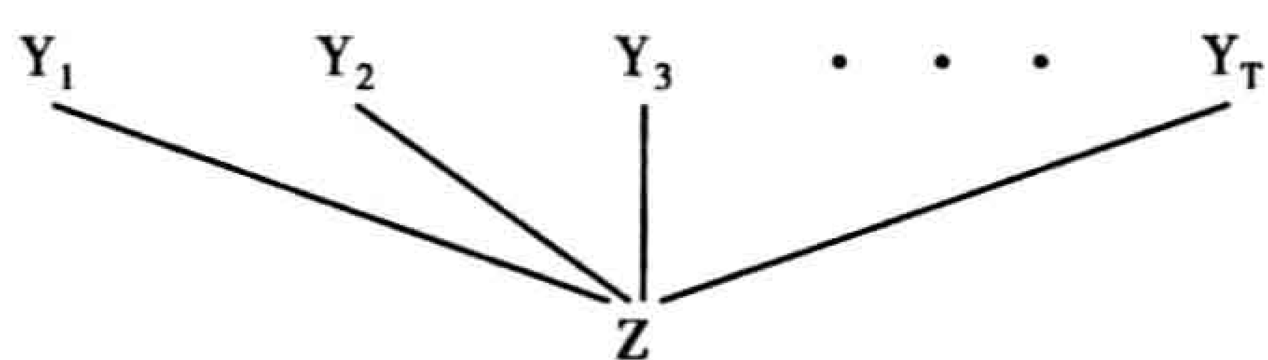


图 13.1　潜类模型的关联图

有时,当所观察到的变量是关于某一概念的几个指标时,比如偏见、虔诚或者对某问题的看法,可以考虑潜类模型。表 10.13 就是这样一个例子,其中,调查对象给出对各种情形下流产是否应当合法的观点。这时,可能存在一个潜变量描述了人们关于合法流产的基本态度,并且在给定这个潜变量的取值后,所观察到的变量之间条件独立。举例来说,这个潜变量可能是一个包括三个类别的分类变量:一类是那些任何情形下都反对合法流产的人,一类是那些总是支持合法流产的人,还有一类依特定的场合对合法流产持不同的态度。

已观察到由$(Y_1,\cdots,Y_T)$进行交叉划分所形成的 T 维列联表,将其与潜变量进行交叉划分的$(T+1)$维表格则是未观察到的。用 I 表示每个 Y_t 所包括的类别数,q 表示 Z 的潜类数量。在所观察到的表格中,令 $\pi_{y_1,\cdots,y_T}=P(Y_1=y_1,\cdots,Y_T=y_T)$。模型假定表格的 I^T 个单元格服从一个多项分布,对于任一给定的单元格,

$$\pi_{y_1,\dots,y_T} = \sum_{z=1}^{q} P(Y_1 = y_1,\cdots,Y_T = y_T \mid Z = z)P(Z = z)。$$

根据潜类模型的条件独立性假定,则有

$$\pi_{y_1,\cdots,y_T} = \sum_{z=1}^{q}\left[\prod_{t=1}^{T} P(Y_t = y_t \mid Z = z)\right]P(Z = z)。\tag{13.1}$$

这是一个关于 I^T 个多项分布概率的非线性模型。

13.1.1　潜类模型的拟合

用$\{n_{y_1,\dots,y_T}\}$表示所观察到的表格中的计数。对该表的 I^T 个单元格求和,多项分布对数似然函数的核函数为

$$\sum n_{y_1,\dots,y_T}\log \pi_{y_1,\dots,y_T}。\tag{13.2}$$

将式 13.1 中相应参数代入上式,通过 Newton-Raphson 算法(Haberman, 1979, Chap. 10)或 EM 算法(Goodman, 1974)可以针对这些参数来最大化函数(13.2)。值得指出的是,潜类模型表明,未观察到的表格满足由符号$(Y_1Z,Y_2Z,\cdots,Y_TZ)$表示的对数线性模型。该模型并未对$\{Y_tZ\}$的关联做任何假定,但是它假定在 Z 的每个类别内$\{Y_t\}$相互独立。

EM 算法的每次迭代过程由两步组成。其中,第 s 次迭代的 E(期望)步利用$\{n_{y_1,\dots,y_T}\}$以及下文将要介绍的关于$(Z\mid Y_1,\dots,Y_T)$的操作性条件分布,来计算未观察到表格中的虚拟计数(pseudo-counts)$\{n^{(s)}_{y_1,\dots,y_T,z}\}$。$M$(最大化)步将$\{n^{(s)}_{y_1,\dots,y_T,z}\}$视为数据本身,并利用迭代再加权最小二乘法或 IPF 等算法来拟合模型(即,对数线性模型$(Y_1Z,Y_2Z,\cdots,Y_TZ)$)。接下来,再利用对未观察到表格的拟合值$\{\mu^{(s)}_{y_1,\cdots,y_T,z}\}$去设定新一轮迭代的 E 步中对应于$\{n_{y_1,\dots,y_T}\}$的$(Z\mid Y_1,\dots,Y_T)$的新的操作性条件分布。这种方法按拟合结果的比例将所观察到的数据分配给未观察到的单元格的虚拟计数中,具体做法为:

$$n^{(s+1)}_{y_1,\dots,y_T,z} = n_{y_1,\dots,y_T}\frac{\mu^{(s)}_{y_1,\dots,y_T,z}}{\sum_{k=1}^{q}\mu^{(s)}_{y_1,\dots,y_T,k}}。$$

这便是在第$(s+1)$次迭代时输入的未观察到的表格的单元格计数。在第$(s+1)$次迭代的 M 步中,它们被用作虚拟数据(pseudo-data)。

最终,EM 算法在未观察到的表格的拟合值处收敛,这些值使得未观察到的表格的拟

合概率在每个潜类中满足相互独立性，并且在所观察到的表格（即，对所有潜类进行求和）中的相应概率使似然函数（式 13.2）最大化。未观察到的表格的拟合概率是对 $(Y_1,\cdots,Y_T,Z)$ 的联合分布的估计。我们可以利用这些拟合概率，计算潜类模型参数 $\{P(Y_t=y_t|Z=z)\}$ 和 $\{P(Z=z)\}$ 的最大似然估计值。

EM 算法在运算上较为简便而且相对稳定。每次迭代都会使似然值上升。但是，它的收敛可能会很慢。有关综述，可参见：Laird（1998）。潜类模型的对数似然函数可能存在局部极大值（local maxima）。因此，不论使用 Newton-Raphson 算法还是 EM 算法，最好利用不同的初始参数值多进行几次拟合。相对而言，EM 算法受选取的初始值的影响更小一些。因而，一些软件在拟合过程中首先使用 EM 算法，然后在拟合过程接近最大似然估计时再转用 Newton-Raphson 算法以加快速度。随着 q 的增加，存在多个局部极大值的可能性也会变大，这时，模型缺乏可识别性的危险也随之上升。

模型参数估计的标准误可以通过对模型所估计的信息矩阵求逆矩阵来获得。在 Newton-Raphson 算法中，这些信息会自动给出，但 EM 算法则不会。在使用 EM 算法时，获得标准误的一种方法是对所观察到的信息矩阵应用 Louis（1982）提出来的一个重要公式。它等于，对未观察到的表格拟合对数线性模型时所观察到的信息矩阵的期望值减去给定观察到的数据后关于 Z 的条件分布的信息矩阵的期望值。Baker（1992）以及 Lang（1992）介绍了有关结果。

通过比较单元格计数的观察值与模型拟合值的卡方统计量，可以对模型进行拟合优度检验。检验的残差自由度 $df=I^T-qT(I-1)-q$。这是因为，多项分布模型（式 13.1）利用在 z 和 t 的每个组合处（共 qT 个组合）的 $(I-1)$ 个参数 $\{P(Y_t=y_t|Z=z),y_t=1,\cdots,I-1\}$ 和另外 $q-1$ 个参数 $\{P(Z=z)\}$ 来描述 I^T-1 个多项分布概率。通常来说，我们会根据变量本身的特点确定 q 的取值，它一般都很小（2 到 4 之间）。如果无法确定，常见的做法是以 $q=2$ 作为起点；如果模型拟合得不够充分，每次对 q 增加 1，直到对拟合结果满意为止。利用专门的软件（附录 A），可以拟合这样的模型。

13.1.2 关于评定者一致性的潜类模型

表 13.1 对第 10.5 节所介绍的一个例子对应的数据进行了扩展。7 位病理学家对 118 张有关子宫颈癌严重程度的图片分别进行了评定。在对评定者之间的一致性进行模型分析时，潜类模型的条件独立性假定往往是合理的。在评定者互不知情的研究设计中，不同病理学家对某个对象的评分之间是相互独立的。如果真实情况属于同一类别的对象之间相对同质，那么在给定真实情况的类别后，不同病理学家的评定结果之间应当接近于相互独立。因此，我们可以设定一个 $q=2$ 个类别的潜类模型，一类针对真实结果为阳性的对象，另一类针对真实结果为阴性的对象。这个模型将由 7 个评定结果所形成的 2^7 联合分布表示为两个 2^7 联合分布的混合，每个真实类别各对应一个。

表 13.1 关于子宫颈癌的评定情况与潜类模型的拟合结果[a]

病理学家								拟合结果		
A	B	C	D	E	F	G	计　数	$q=1$	$q=2$	$q=3$
0	0	0	0	0	0	0	34	1.1	23.0	33.8
0	0	0	0	1	0	0	2	1.6	6.6	2.0
0	1	0	0	0	0	0	6	2.2	12.7	6.3
0	1	0	0	0	0	1	1	2.8	1.7	1.5
0	1	0	0	1	0	0	4	3.3	3.6	3.0
0	1	0	0	1	0	1	5	4.2	0.5	4.7

续表

病理学家							计 数	拟合结果		
A	B	C	D	E	F	G		$q=1$	$q=2$	$q=3$
1	0	0	0	0	0	0	2	1.4	3.0	2.1
1	0	1	0	1	0	1	1	1.6	0.2	0.2
1	1	0	0	0	0	0	2	2.8	1.7	1.3
1	1	0	0	0	0	1	1	3.5	0.3	1.6
1	1	0	0	1	0	0	2	4.2	0.5	2.9
1	1	0	0	1	0	1	7	5.3	3.7	6.5
1	1	0	0	1	1	1	1	1.4	2.6	1.4
1	1	0	1	0	0	1	1	1.3	0.1	0.1
1	1	0	1	1	0	1	2	2.0	4.3	2.6
1	1	0	1	1	1	1	3	0.5	3.1	2.0
1	1	1	0	1	0	1	13	3.3	11.5	9.6
1	1	1	0	1	1	1	5	0.9	8.4	8.7
1	1	1	1	1	0	1	10	1.2	13.5	13.6
1	1	1	1	1	1	1	16	0.3	9.9	12.3

a 模型拟合结果由 Latent Gold 软件(Statistical lnnovations, Belmont MA)给出。1,是;0,否。
来源:数据基于:Landis and Koch(1977),未显示空单元格。

表 13.2 给出了几个潜类模型(其中包括一个混合模型,将在第 13.2.4 节介绍)的拟合结果。由于所观察到的表格存在数据稀疏问题,偏离度主要用来进行模型间的比较。但是,这是一种非正式的模型比较,因为在对比具有不同潜类数量的模型的偏离度时,卡方分布并不适用。包括 q 个潜类的模型是包括 $q^* > q$ 个潜类的模型的一个特例,在后者中令对 $z > q$ 的潜类都满足 $P(Z=z)=0$ 便得到前者,相应地,这些参数落在了参数空间的边界上。而普通的卡方似然比检验要求参数必须落在参数空间的里面(即,对于 $z=1,\cdots,q^*$,都有 $0<P(Z=z)<1$)。

表 13.1 还给出了当 $q=1,2,3$ 时,潜类模型对非空单元格的拟合值(每个空单元格同样都对应着一个拟合值,但这里没有给出)。具有 $q=1$ 个潜类的模型实际就是要求 7 个评定者的评定结果相互独立的模型,它等价于对数线性模型($Y_1,Y_2,\cdots,Y_7$)。不出所料,该模型的拟合结果很差。当 $q=2$ 时,仍存在明显证据表明,模型拟合得不充分。例如,所拟合的每个病理学家评定结果为阴性的计数为 23.0,而相应的观察计数则为 34(在表 13.2 中,这个模型所对应的 G^2 值较小并不表示它对数据拟合得很好;在第 9.8.4 节中我们已经指出,当大多数拟合值都很接近于 0 时,G^2 统计量一般非常保守)。当 $q=3$ 时,潜类模型对数据的拟合似乎是充分的。

表 13.2 对表 13.1 数据所拟合的潜类模型的似然比统计量[a]

潜类的数量	模 型	偏离度(G^2)统计量	df
1	相互独立性	476.8	120
2	潜类	62.4	112
	Rasch 混合	67.6	118
3	潜类	15.3	104
	Rasch 混合	27.5	116
4	潜类	6.4	96
	Rasch 混合(准对称性)	23.7	114

a 模型由 Latent Gold 软件(Statistical lnnovations, Belmont MA)拟合。

以给定的潜类 z 为条件,分析每个病理学家诊断结果的估计概率 $P(Y_t=1 \mid Z=z)$ 有助于我们理解这些潜类的含义。表 13.3 给出了三个潜类模型的相应估计概率。这些结果表明:(1)第一个潜类是所有病理学家(除了 B 有时不一致)都认为没有癌症的情况;(2)第三个潜类是 A,B,E 和 G 都认为存在癌症,而且 C 和 D 也基本上同意的情况;(3)第二个潜类指的是出现了很强的不一致的情况,其中 C,D 和 F 很少诊断为癌症,而 B,E 和 G 却往往都诊断为癌症。模型所估计的落在三个潜类的比例分别为 $\hat{P}(Z=1)=0.37$,$\hat{P}(Z=2)=0.18$,以及 $\hat{P}(Z=3)=0.45$。根据该模型,有 18% 的案例属于结果存疑的潜类。

表 13.3 在包括 3 个潜类的潜类模型和 Rasch 混合模型中诊断为癌症的估计概率[a]

模 型	潜 类	病理学家						
		A	B	C	D	E	F	G
潜类	1	0.057	0.138	0.000	0.000	0.055	0.000	0.000
	2	0.513	1.00	0.000	0.058	0.751	0.000	0.631
	3	1.000	0.981	0.858	0.586	1.000	0.476	1.000
Rasch	1	0.022	0.150	0.001	0.000	0.047	0.000	0.022
混合	2	0.611	0.923	0.052	0.015	0.774	0.009	0.611
	3	0.994	0.999	0.853	0.617	0.997	0.483	0.994

a 模型由 Latent Gold 软件(Statistical Innovations, Belmont, MA)拟合。

使用潜类模型的一个危险在于,在解释结果时过分解读潜变量的表面含义,这个问题对于有关连续结果变量的因子分析同样存在。在上文的例子中,很容易将潜类 1(潜类 3)当作确实不存在癌症(患有癌症)的情况。因而,当给定某个对象落在潜类 1(潜类 3)时,很容易将没有癌症(有癌症)的评定结果认为是必然正确的判断。大家应当认识到潜变量的尝试性特征,避免犯具体化(reification)的错误,即把一个抽象的含义误认为客观的存在(Gould, 1981)。

利用模型的参数估计以及贝叶斯定理,我们也可以估计 $P(Z=z \mid Y_t=y_t)$ 和 $P(Z=z \mid Y_1=y_1,\cdots,Y_T=y_T)$。例如,如果一个病理学家评定为"是",那么这个对象落在评定结果一致认为是阳性潜类的估计概率是多少?我们先介绍一个较为简单的模型,然后将在第 13.2.5 节进一步分析这个问题。

Espeland 和 Handelman(1989)、Uebersax(1993)、Uebersax 和 Grove(1990,1993),以及 Yang 和 Becker(1997)介绍了有关评定者一致性与诊断结果精确性的各种潜变量模型。我们也可以运用第 11 章和第 12 章中介绍的方法,比如具有连续潜变量而不是分类潜变量的模型。例如,利用 logistic-正态随机截距模型可以在不同 t 之间进行关于 $P(Y_t=1)$ 的对象别比较。

13.1.3 关于捕获-再捕获数据的潜类模型

接下来,我们利用潜类模型来对估计总体规模的捕获-再捕获数据进行分析。在第 12.3.6 节中,我们介绍过如何使用 logistic-正态广义线性混合模型来处理这类问题。在总抽样次数为 T 的情况下,数据可以通过一个 2^T 列联表来表示,其中每次抽样的结果类别均为(被捕获,未被捕获)。这时,对总体规模的预测相当于对表中缺失的单元格计数的预测,这个缺失的单元格计数表示在各次抽样中均未被捕获的数量,将其与其他单元

格计数加总就可以得到总体规模。

在包括两个潜类的潜类模型中，总体被看作是两种不同类型的子总体的混合体，这些类型可能取决于基因或环境因素。在同一类型内，被捕获概率具有同质性，但是具体到某个对象属于哪一类型是未知的。这个模型反映了一种对相互独立性模型与 logistic-正态模型的折中，前者假定只存在一个潜类即具有完全同质性，后者则假定存在一个连续的被捕获概率的分布而不仅仅是两个潜类。

我们通过表 12.6 中有关捕获雪兔的数据来加以说明，这里 $T=6$。相互独立性模型所预测的结果为 $\hat{N}=75$，它的 95% 的剖面似然置信区间是(70,83)。包括两个潜类的潜类模型预测 $\hat{N}=85$，相应的剖面似然置信区间为(74,106)。包括三个潜类的潜类模型的结果与此相似。与第 12.3.6 节中的 logistic-正态模型所给出的区间(75,154)相比，上述这些区间似乎都太小了。究其原因，简单的潜类模型可能无法描述个体之间的所有异质性。这时，假定存在一个连续的潜变量比只包括几个潜类的离散变量更为合理。我们将在下一节利用相应模型继续分析这些数据。

13.2 非参数随机效应模型

在普通的广义线性混合模型中，尽管关于随机效应服从正态分布的假定很常见，也很有吸引力，但是我们很少能够对此进行细致的检查。例如，在研究正态广义线性混合模型时，Verbeke 和 Lesaffre(1996)指出，在随机效应服从正态分布的假定下，即使真实值的分布与正态分布相去甚远，它们的预测值常常也服从正态分布。由此带来的一个突出问题是，模型的错误设定可能会带来负面后果，这一点对有关随机效应的任意参数化假定都适用。为了对这种假定的灵敏度进行检验，我们可以利用随机效应分布的其他假定或更一般的假定来拟合广义线性混合模型。

13.2.1 未设定随机效应分布的 Logit 模型

非参数方法(如：Aitkin, 1999)可以避免模型设定错误可能导致的不良后果。在该方法中，我们对一个有限集(finite set)的质点(mass points)应用未设定的随机效应分布。这些质点的位置以及它们的概率构成模型的参数。质点的数量可以是固定的；当质点数量本身也未知时，我们在拟合过程中将其视为给定的，然后逐步增加它的值，直到似然函数达到最大值为止。通常来说，最大化过程只需要相对较少的质点数量，即使允许存在连续的混合分布，对相应分布的非参数估计也只需要有限数量的点(如：Lindsay et al., 1991)。事实上，对一个仅包括两个质点的模型进行拟合所得到的固定效应估计值，往往与完全最大化的结果非常相似。只有不特别关注随机效应的分布本身时，非参数方法才有意义，这是因为即便样本规模非常大，关于随机效应分布的非参数估计一般仍然很差。

非参数模型的拟合实际上比具有正态随机效应的模型更为简单，因为决定似然函数的积分简化为一个有限项的求和。在第 13.2.4 节中，我们将通过 Rasch 模型对此进行具体讨论。可以通过专门的软件来拟合非参数混合模型(附录 A)。然而，非参数方法也有它的缺点。例如，当模型中包括多元随机效应时，它无法像正态分布模型那样给出随机效应之间的简单的相关结构。当模型的质点数量不同时，不能通过标准的推断方法进行模型比较，因为此时一个模型落在另一个模型的参数空间的边界上。另外，关于随机效

应分布的最大似然估计通常会对等于 ± ∞ 的值赋予一定的权数，尽管在分析二分数据时这对于识别一个群组中所有观测值的结果变量估计概率都等于 1 或 0 的子样本是有价值的，但是它却无法描述所估计的方差结构的异质性。

我们重新分析表 10.13 中有关合法流产态度的数据，用以展示非参数方法。在第 12.3.2 节中，我们对相应数据拟合了 logistic-正态模型（式 12.10）

$$\text{Logit}\left[P(Y_{it}=1 \mid u_i)\right] = u_i + \beta_t + \gamma x, \tag{13.3}$$

其中 x = 性别，参数 $\{\beta_t\}$ 代表合法流产的三种可能情况。在这里，我们将 u_i 视为非参数的，相应的似然函数在一个两点混合分布下进行最大化。所估计的有关流产态度问题的效应为 $\hat{\beta}_1-\hat{\beta}_3=0.83$（SE = 0.16），$\hat{\beta}_2-\hat{\beta}_3=0.30$（SE = 0.16），以及 $\hat{\beta}_1-\hat{\beta}_2=0.52$（SE = 0.16）。这些结果与表 12.3 所给出的正态随机效应模型的结果（第 12.3.2 节）相似。

13.2.2 非参数的混合 Logistic 回归

Follman 和 Lambert（1989）介绍了一个质点数量事先设定的例子。他们分析一种毒药的剂量对某一属的原生动物死亡概率的影响。表 13.4 给出了有关数据。他们假定该属包括两种潜在的类型。

表 13.4 按照不同剂量水平划分的原生动物数量与死亡数量

毒药剂量	实验数量	死亡数量	毒药剂量	实验数量	死亡数量
4.7	55	0	5.1	53	22
4.8	49	8	5.2	53	37
4.9	60	18	5.3	51	47
5.0	55	18	5.4	50	50

来源：Follman and Lambert（1989）。经《美国统计学会期刊》（the *Journal of the American Statistical Association*）授权重印。

令 $\pi_i(x)$ 表示该属中的第 i 类动物在剂量水平的对数为 x 时的死亡概率，$i=1,2$。令 ρ 表示原生动物隶属于该属的第 1 类的概率。他们的模型设定

$$\pi(x) = \rho\pi_1(x) + (1-\rho)\pi_2(x), \text{其中 Logit}\left[\pi_i(x)\right] = \alpha_i + \beta x,$$

这里 ρ 是未知参数。关于 $\pi(x)$ 的曲线是对两条形状相同、截距不同的曲线的加权平均。

普通的 logistic 回归模型是上述模型中当 $\rho=1$ 时的特例。它的拟合结果为 $\text{Logit}[\hat{\pi}(x)] = -68.4+42.1x$（$\hat{\beta}=42.1$ 的标准误为 SE = 3.8），这个结果很差，其偏离度 $G^2=24.7$（df = 6）。混合模型的拟合结果为

$$\hat{\pi}(x) = 0.34\hat{\pi}_1(x) + 0.66\hat{\pi}_2(x),$$

其中

$$\text{Logit}\left[\hat{\pi}_1(x)\right] = -196.2+124.8x, \quad \text{Logit}\left[\hat{\pi}_2(x)\right] = -205.7+124.8x,$$

关于 $\hat{\beta}=124.8$ 的标准误为 SE = 25.2。图 13.2 显示了相应的拟合结果。这一拟合结果比上面的模型好得多，$G^2=3.4$（df = 4），也即，通过加入两个参数——一个额外的截距和一个混合概率，二倍的最大对数似然值之差等于 24.7 − 3.4 = 21.3。Follman 和 Lambert 指出，当具有八种不同的剂量水平时，对这个模型最多只有两个混合点是可识别的。

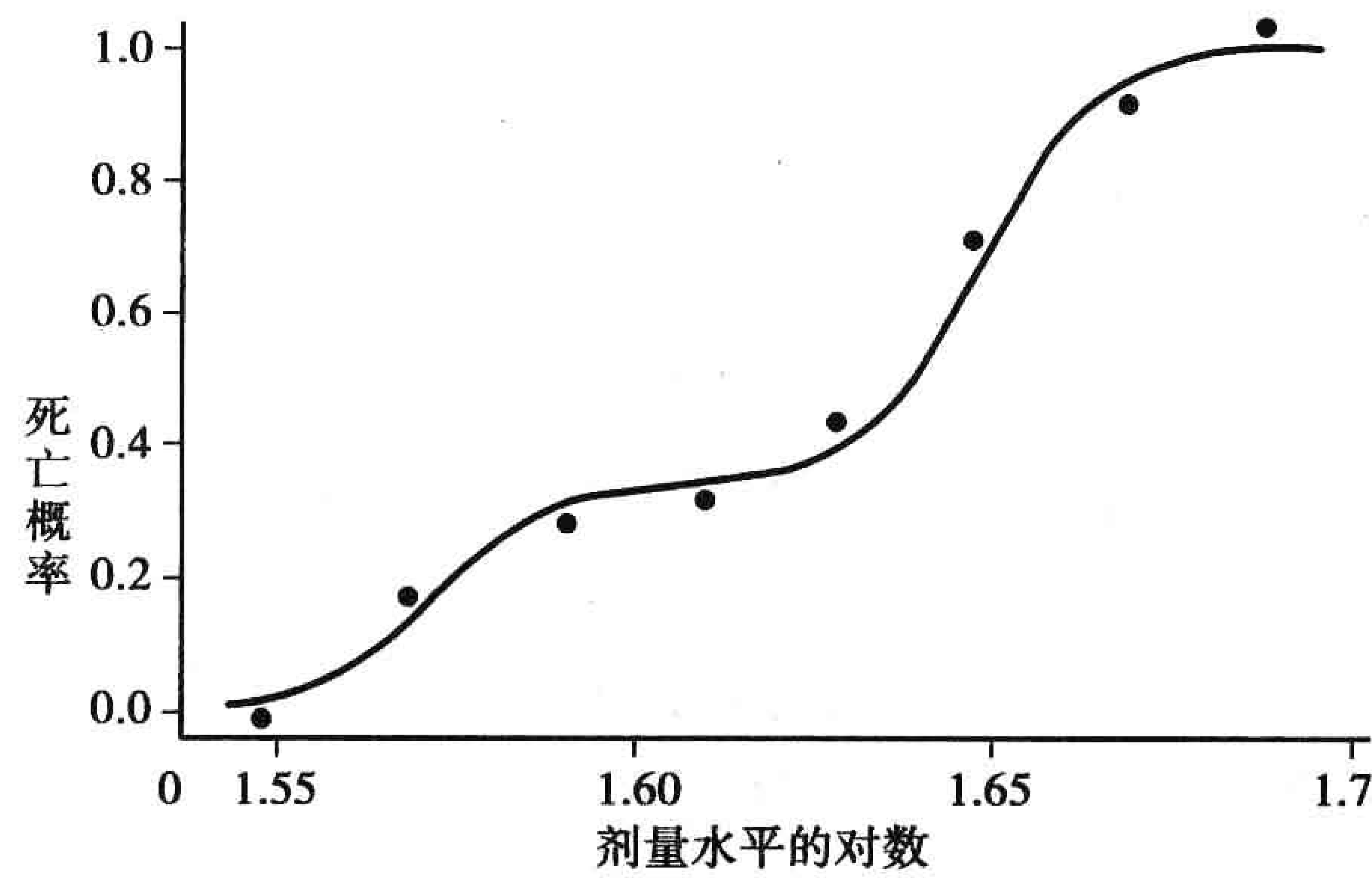


图 13.2 对表 13.4 数据拟合二分混合 logistic 回归的结果

(模型由 Latent Gold(Statistical Innovations, Belmont, MA)拟合)

普通的广义线性混合模型假定一组 logistic 曲线的混合服从正态分布。与 $\rho=1$ 时的普通 logistic 模型相比,它所对应的偏离度仅下降了 1.7。

13.2.3 模型设定错误的问题严重吗?

在运用广义线性混合模型时,有必要花费精力去考虑有关随机效应正态假定以外的其他选择吗,无论是参数形式还是非参数形式? 到目前为止,考察模型设定错误的研究还比较少见。对于 logistic 随机截距模型,在随机效应分布的不同假定下所得出的回归系数的估计往往很相似。因而,选取不正确的随机效应分布一般不会对这些效应估计造成偏误。当随机效应的真实分布呈偏态(skewed)时,有关正态截距的估计值会出现一定的偏误(Neuhaus et al., 1992)。通常情况下,对随机效应分布的选取基本不会影响估计的效率。

当真实的随机效应分布与正态分布相去甚远时,logistic-正态估计值会损失一定的效率。当真实分布服从一个方差构成很大的两点混合分布时,就会出现这种情况。笔者曾与 B. Caffo 就多种模型探讨过这一问题,包括简单的单维随机效应模型。在第 i 个群组中,令 y_{it}表示一个伯努利变量,它满足

$$\text{Logit}\left[P(Y_{it}=1\mid u_i)\right]=\alpha+u_i, i=1,\cdots,n, t=1,\cdots,T, \tag{13.4}$$

其中,$\text{var}(u_i)=\sigma^2$。对 n, T, α 和 σ 赋不同的值,并尝试关于 u_i 的各种真实分布,包括正态分布、均匀分布、指数分布以及二项分布,我们对模型结果进行了模拟。通常来说,当真实分布不服从正态分布时,正态分布假定不会产生负面影响。同时,当真实分布服从正态分布时,运用非参数方法也不会导致很大的效率损失(Neuhaus 和 Lesperance(1996)在分析有关模型时指出了这一点)。然而,当真实分布是一个两点混合分布时,随着 α 和 T 的取值的增加,正态分布方法在估计$\{\mu_i=P(Y_{it}=1\mid u_i)\}$时会造成效率损失。例如,当 $n=T=30$ 且 $\alpha=0$ 时,该混合分布在每个点的概率都是 0.5;当 $\sigma=0.5$ 时,(正态,非参数)方法下关于$|\hat{\mu}_i-\mu_i|$的期望值为(0.06,0.05);当 $\sigma=1.0$ 时,相应期望值为(0.06,0.02);当 $\sigma=2.0$ 时,相应期望值为(0.04,0.01)。两种方法关于 α 的估计之间的差异相对更小一些。

第 13.2.2 节所讨论的例子取自 Follman 和 Lambert(1989),其中模型包括一个协变量但是 $T=1$。该例展示了运用 logistic-正态广义线性混合模型所可能导致的效率损失。

运用两点混合模型的拟合结果为 $\hat{\beta}=124.8$，SE $=25.2$，其中 $\hat{\beta}/\text{SE}=4.9$。而在正态混合模型中，$\hat{\beta}=65.5$，SE $=19.5$，其中 $\hat{\beta}/\text{SE}=3.4$。

我们的研究表明，除非随机效应的分布极端偏离正态分布，正态广义线性混合模型一般不会导致估计偏误或效率损失。但是，Heagerty 和 Zeger（2000）（另见：McCulloch，1997）指出，其他形式的模型设定错误会造成更严重的后果。他们认为，与相应的边际模型相比，在估计随机效应模型中的回归参数时，估计偏误对不同随机效应假定的灵敏度更大。他们利用一个随机效应的方差取决于协变量取值的模型对此进行了说明。他们的结论是，组间效应可能比组内效应对随机效应分布的正确设定更敏感。因而，在估计组间效应时，使用边际模型具有不受分布设定错误影响的优点。

13.2.4 Rasch 混合模型

根据第 12.1.4 节，在结果变量为二分变量的情况下，关于第 i 个对象的第 t 个项目的 Rasch 模型是

$$\text{Logit}\left[P(Y_{it}=1\mid u_i)\right]=u_i+\beta_t, t=1,\cdots,T。\qquad(13.5)$$

广义线性混合模型将 $\{u_i\}$ 视为服从正态分布的随机效应。Lindsay 等（1991）研究了当该模型中的 u_i 只能取 q 个有限数量值的情况。对所有 i，将潜变量 u_i 的分布表示为

$$P(U=a_k)=\rho_k, k=1,\cdots,q,$$

其中，$\{a_k\}$ 和 $\{\rho_k\}$ 未知。出于模型可识别性的考虑，可以对这个分布加以限定，如 $\sum_k \rho_k a_k=0$，或者对 $\{\beta_t\}$ 加以限定。这个模型被称为 Rasch 混合模型（*Rasch mixture model*）。

与其他随机效应模型相同，Rasch 混合模型是一个潜变量模型。随机效应 u_i 是未观测到的，并且在 u_i 的每个给定取值处假定 T 个结果变量条件独立。它与有关二分结果变量的包括 q 个潜类的普通潜类模型（第 13.1 节）不同，因为它假定 $P(Y_{it}=1\mid u_i)$ 满足式 13.5 的结构，而式 13.1 潜类模型并未假定 $P(Y_t=y_t\mid Z=z)$ 的结构。

与具有正态随机效应的广义线性混合模型相比，这个模型的拟合更为简便，因为它用一个有限项的求和替代了广义线性混合模型中决定似然函数的复杂积分。对应于此模型，回答序列 $(y_1,\cdots,y_T)$ 的边际概率为

$$\pi_{y_1,\ldots,y_T}=\sum_{k=1}^{q}\rho_k\left[\prod_{t=1}^{T}\frac{\exp[y_t(a_k+\beta_t)]}{1+\exp(a_k+\beta_t)}\right]。$$

将其代入多项分布对数似然函数（式 13.2），可以通过 Newton-Raphson 算法或 EM 算法获得关于 $\{a_k,\rho_k\}$ 和 $\{\beta_t\}$ 的最大似然估计。随着 q 的增加，最大似然值上升，拟合结果进一步改善。但是，Lindsay 等（1991）表明，在存在 T 个项目的情况下，一旦 q 值增加到 $q=(T+1)/2$ 后，似然值就不再变动。这时，该模型对所观察到的 2^T 表格的拟合结果与准对称性模型（式 10.33）相同。因此，这个较简单的潜类模型在所观察到的变量之间具有对称的条件关联结构。Arminger 等（2000）对 Rasch 混合模型进行了扩展，用以包括协变量的情况。

13.2.5 关于评定者一致性的模型分析

对于 7 位病理学家评定癌症的数据（表 13.1），表 13.2 中也给出了 Rasch 混合模型的拟合结果。在这里，式 13.5 模型中的 $P(Y_{it}=1\mid u_i)$ 表示第 t 位病理学家关于第 i 张图

片的诊断结果为癌症的概率。当 $q=3$(即 u_i 可以取 3 个值)时,该模型的拟合结果并不比潜类模型差。由于总共有 $T=7$ 个评定者,这里的离散混合分布最多可以包括 $(T+1)/2=4$ 个点。当 $q=4$ 时,该模型等价于准对称性模型,其拟合结果并不明显好于 $q=3$ 的模型。

图13.3显示了当 $q=3$ 时的Rasch混合模型的估计值 $\{\hat{\beta}_t\}$,其中模型设定 $\sum_t \hat{\beta}_t = 0$。这些估计值描述了在潜变量的每个取值水平上,病理学家的评定结果分布之间的变动性。例如,对于一个给定的潜类,所估计的病理学家 B 诊断结果为癌症的发生比是病理学家 A 的相应发生比的 $\exp(3.52-1.48)=7.7$ 倍。病理学家B最倾向于做出癌症的诊断,而D和F最不倾向于做出癌症的诊断。基于成对差异 $\hat{\beta}_t-\hat{\beta}_s$ 的标准误,该图还给出了关于病理学家两两之间的 21 个配对的 90% 的 Bonferroni 比较(Bonferroni comparison)的结果。

病理学家	F	D	C	A	G	E	B
估计值	−3.70	−3.15	−1.87	1.48	1.48	2.26	3.52

比较:F–D、D–C 各有一条连线;A–G–E 一条连线;E–B 一条连线。

图 13.3 Rasch 混合模型对每位病理学家的估计结果以及 90% 的 Bonferroni 同步比较结果(Bonferroni simultaneous comparison)

以某张图片的潜变量取值水平 k 为条件,对于第 t 位病理学家,

$$\exp(\hat{a}_k+\hat{\beta}_t)/\left[1+\exp(\hat{a}_k+\hat{\beta}_t)\right]$$

是所估计的诊断结果为癌症的概率。表 13.3 报告了这些估计结果,其中 $\hat{a}_1=-5.25$,$\hat{a}_2=-1.02$,$\hat{a}_3=3.63$。这些结果与普通潜类模型的估计结果相似,相对更为平滑,落在边界上的估计值相对较少。同样,当潜变量取值为 1 时,病理学家一般诊断为没有癌症,当取值为 2 时诊断结果存在很大的不一致,而当取值为 3 时他们倾向于诊断为癌症。模型所估计的潜类的比例为 $\hat{\rho}_1=0.37$,$\hat{\rho}_2=0.19$,$\hat{\rho}_3=0.43$,即有 19% 的情况落在诊断结果不一致这一类。

式 13.5 模型表明,对于 U 的取值水平 k 和 ℓ,每个 Y_t 与 U 之间的关联具有对数发生比之比 (a_k-a_ℓ)。例如,在第三个潜类中所估计的病理学家诊断为癌症的发生比是第一个潜类的相应发生比的 $\exp[3.63-(-5.25)]>7\,000$ 倍。按照表 13.3 中所估计的概率,病理学家 A 对应的 (a_k-a_ℓ) 等于 $\exp[(0.994/0.006)/(0.022/0.978)]$。$\{\hat{a}_k-\hat{a}_\ell\}$ 的值很大表明在每位病理学家的评定结果与潜变量之间存在强关联。由此推出,在每对病理学家的评定结果之间也存在强关联(模型拟合的关于每对评定者之间的发生比之比的值大约在 7 到 400 之间变动)。然而,$\{\hat{\beta}_t\}$ 之间的巨大差异表明,这七次评定中存在明显的边际异质性,这导致了评定结果之间两两对比的一致性程度各异。

相互独立性模型是 Rasch 混合模型在 $q=1$ 时的一个特例,也即,$\rho_1=1$。对于表 13.1,具有 $q=3$ 的 Rasch 混合模型只比相互独立性模型多四个参数(即 ρ_k 和 a_k,$k=1,2$)。但是,它对数据拟合得很好,而且便于解释。对此模型的更详细讨论以及通过设定 $a_1-a_2=a_2-a_3$ 以简化模型的做法,参见:Agresti and Lang(1993b)。

13.2.6 关于捕获-再捕获数据的其他模型

在第 13.1.3 节中,我们利用潜类模型对捕获-再捕获实验数据进行了分析。除此之

外,有关分析也可以考虑使用 Rasch 混合模型。具有两个潜类的式 13.5 模型的拟合结果为 $\hat{N}=77$,相应的 95% 的剖面似然置信区间为(71,87)。这个区间似乎过小而缺乏可信性。因而,更现实的做法是允许被捕获概率服从连续分布。将式 13.5 模型中的 u_i 视为正态分布而不是二项分布就可以解决这一问题,在第 12.3.6 节中我们已经运用相应模型分析了这些数据。

那么,除了参数的随机效应模型之外,我们还可以考虑什么模型呢?一种可能是对数线性模型(Cormack, 1989)。对数线性模型是边际模型,对应的是将所有对象求平均的概率。令 Y_t 表示在第 t 次抽样中一个随机选取的对象对应的二分结果变量,其类别为(被捕获,未被捕获)。这时,最简单的模型假定每次捕获的发生是相互独立的,由(Y_1, Y_2,…,Y_T)来表示,它等价于 $\sigma=0$ 的 logistic-正态模型(式 13.5)以及 $q=1$ 的潜类模型(式 13.1)。允许每对表示捕获结果的变量之间存在关联的模型相对更为合理,它等价于对数线性模型(Y_1Y_2, Y_1Y_3, …, $Y_{T-1}Y_T$)。另外,具有马尔科夫结构的模型,如(Y_1Y_2, Y_2Y_3,…,$Y_{T-1}Y_T$),在此可能也是有价值的。通常情况下,这类数据往往不足以拟合过于复杂的对数线性模型。在以上任意模型中,模型通过对 2^T-1 个所观察到的单元格的拟合来预测剩下单元格中未被观察到的计数。

非参数随机效应模型与对数线性模型之间存在一定的联系。在第 13.2.7 节中我们将介绍,假定 u_i 具有非参数形式的式 13.5 模型在边际上对应于准对称性形式的对数线性模型。然而,式 10.33 准对称性模型本身对分析捕获-再捕获问题没有帮助,因为在缺失的单元格中放入任意计数都与该模型一致。这一模型包括一个专门针对该单元格的交互项参数,其似然方程设定该单元格计数等于拟合值,因而,其他单元格中的信息对估计该单元格的期望频数没有意义。但是,一些特殊情况的准对称性模型是有用处的(Darroch et al., 1993),例如要求每对抽样之间的关联都相等的对数线性模型。与 logistic-正态模型相似,这个满足*可交换的关联*(*exchangeable association*)的模型只比相互独立性模型多一个参数。

对于表 12.6 的数据,具有可交换的二维关联的模型结果为 $\hat{N}=90.5$,相应置信区间为(75,125)。这个区间和 $q=2$ 时 Rasch 混合模型的区间(71,87),都大大小于 logistic-正态模型(第 12.3.6 节)估计的相应区间(75,154)。在捕获-再捕获实验中,关于 N 的估计值 $\hat{N}$ 和置信区间在很大程度上取决于模型的选择。这一问题在预测中无法避免。对 N 的估计需要从所观察到的被捕获了 1,2,…,T 次的对象中去外推一次也未被捕获到的对象的数量。这时,标准的拟合优度指标意义有限。可能两个模型都对数据拟合得很好,但给出的关于未观察到的计数的估计却非常不同。例如,对于雪兔的数据,相互独立性和具有二维关联的对数线性模型对所观察到的单元格拟合得都不错(对于相互独立性模型,$G^2=58.3$,df=56;对于二维关联模型,$G^2=32.4$,df=41),可是,两个模型所估计的 $\hat{N}$ 分别为 75 和 105。

根据简约模型的好处,较简单的模型通常给出的关于 N 的置信区间也较小。但是这对于捕获-再捕获数据来说并不一定就是优点:我们确实偏好关于 N 的一个较小的置信区间,但是不能以严重牺牲实际置信区间的大小为代价。对有关对象不切实际地做同质性假设的区间可能会过于乐观。模拟研究结果表明,即便当模型设定出现微小错误时,区间的实际涵盖概率往往都低于其名义水平。允许对象间存在异质性会导致较大的区间。当存在严重的总体异质性时,我们甚至无法做出有用的结论,因为相应的区间可能会非常大(Burnham and Overton, 1978, Coull and Agresti, 1999)。

13.2.7 非参数混合与准对称性

在式 13.5Rasch 模型中不限定 u_i 的分布时，它在边际上对应于准对称性对数线性模型(Dorroch, 1981;Tjur, 1982)。这一结论在第 10.4.2 节就已经提及，现在我们给出其证明过程。

令 $\mathbf{Y}_i$ 表示关于第 i 个对象的 T 个结果变量的取值组合。对于可能的结果 $\mathbf{y}=(y_1,\cdots,y_T)$，其中每个 $y_t=1$ 或 0：

$$P(\mathbf{Y}_i=\mathbf{y}\mid u_i)=\prod_t\left[\frac{\exp(u_i+\beta_i)}{1-\exp(u_i+\beta_i)}\right]^{y_t}\left[\frac{1}{1+\exp(u_i+\beta_t)}\right]^{1-y_t}$$
$$=\frac{\exp\left[u_i(\sum_t y_t)+\sum_t y_t\beta_t\right]}{\prod_t[1+\exp(u_i+\beta_t)]}。$$

用 F 表示 u_i 的累积分布函数，这样，对于任一随机选取的对象，结果组合 $\mathbf{y}$ 发生的边际概率为(略去关于对象的下标)

$$\pi_{y_1,\cdots,y_T}=E_UP(\mathbf{Y}=\mathbf{y}\mid U)=\exp\left(\sum_t y_t\beta_t\right)\int\frac{\exp\left[u(\sum_t y_t)\right]}{\prod_t[1+\exp(u+\beta_t)]}\mathrm{d}F(u)。$$

这个概率是对数似然函数的一部分，对于 2^T 个单元格的多项分布的可能结果 $\mathbf{y}$ 而言，该函数为式 13.2。无论选取哪个 F 函数，以上积分运算都非常复杂。但是，该函数只通过 $\sum_t y_t$ 依赖于数据。更一般化的模型通过对每个 $\sum_t y_t$ 的取值分别设定参数来替代上面的积分。相应模型具有如下形式：

$$\log\pi_{y_1,\cdots,y_T}=\sum_t y_t\beta_t+\lambda_{y_1+\cdots+y_t}。\qquad(13.6)$$

其中，最后一项表示在每个 $\sum_t y_t$ 取值处的相应参数。

式 13.6 模型对应的边际模型中的交互项不随着结果变量 $\mathbf{y}$ 的排列变化而变化，因为每种排列的求和 $\sum_t y_t$ 都相同。因此，它是一个准对称性对数线性模型(式 10.33)。无论 F 的形式如何，边际模型具有相同的主效应结构，并且它的交互项是式 13.6 模型的交互项的一个特例。因而，通过有关对数线性模型的普通最大似然估计，可以得到关于 $\{\beta_t\}$ 的一致性估计。事实上，Tjur(Tjur, 1982)证明，这些估计值也是将 $\{u_i\}$ 视为固定效应并以其充分统计量为条件的条件最大似然估计。式 13.6 模型中的交互项参数源于结果变量取值之间的相依性，这是由 $\{u_i\}$ 的异质性导致的。

我们以关于合法流产态度的数据为例来加以说明。这些数据曾在第 10.7.2 节和第 12.3.2 节分析过，并在第 13.2.1 节运用非参数随机效应方法进行过分析。对于模型式 13.3，通过拟合一个准对称性对数线性模型，可以估计个体在不同项目间的比较($\beta_t-\beta_s$)。令 $\mu_g(y_1,y_2,y_3)$ 表示性别为 g 的对象对第 t 个项目($t=1,2,3$)的回答为 y_t 的期望频数，其中 $y_t=1$ 表示赞成合法流产，0 表示反对。该对数线性模型为

$$\log\mu_g(y_1,y_2,y_3)=\beta_1y_1+\beta_2y_2+\beta_3y_3+\gamma g+\lambda_{y_1+y_2+y_3}。\qquad(13.7)$$

对于 $y_1+y_2+y_3=k$，λ_k 表示调查对象对三个项目中的 k 个表示支持的所有单元格，$k=0,1,2,3$。模型的最大似然拟合结果为 $G^2=10.2$，df $=9$，并且 $\hat{\beta}_1-\hat{\beta}_2=0.521$(SE $=0.154$)，$\hat{\beta}_1-\hat{\beta}_3=0.828$(SE $=0.160$)，以及 $\hat{\beta}_2-\hat{\beta}_3=0.307$(SE $=0.161$)。这些结果与正态随机效应估计结果(表 12.3)以及第 13.2.1 节中的非参数随机效应估计都很相似。它们也是关于式

13.3 模型的条件最大似然估计,其中$\{u_i\}$被视为给定的。然而,利用这种方法或者条件最大似然法,我们无法估计组间效应,比如式 13.7 模型中的性别效应(式 13.7 模型中的γ参数指的是男性和女性样本的相对规模,它与式 13.3 模型中的性别效应不同)。

13.3 β-二项分布模型

β-二项(beta-binomial)分布模型是除了具有正态随机效应的二分广义线性混合模型之外的另一种参数混合模型。与其他混合模型相同,β-二项分布模型在给定的参数值处也假定服从二项分布,但它的边际分布所允许的变动性比二项分布更大。因此,当普通二项分布模型出现过度离散问题时,β-二项分布模型提供了一种处理此类问题的办法。

13.3.1 β-二项分布

β-二项分布是由多个二项分布混合所得的β分布。该分布假定:(a)给定π,Y服从二项分布 $\text{bin}(n,\pi)$;(b)π服从β分布。

β概率密度函数为

$$f(\pi;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}, 0 \leqslant \pi \leqslant 1, \tag{13.8}$$

其中$\Gamma(\cdot)$为γ(gamma)函数,参数$\alpha>0$且$\beta>0$。令

$$\mu = \frac{\alpha}{\alpha+\beta}, \theta = 1/(\alpha+\beta)。$$

关于π的β分布的均值和方差分别为

$$E(\pi) = \mu, \text{var}(\pi) = \mu(1-\mu)\theta/(1+\theta)。$$

当α和β都大于 1.0 时,该分布是单峰的:如果$\alpha<\beta$,它向右偏斜;如果$\alpha>\beta$,它向左偏斜;如果$\alpha=\beta$,它是对称的。当$\alpha=\beta=1$时,β分布简化为均匀分布。

对π的β分布求平均可得,Y在边际上服从*β-二项分布*(*beta-binomial distribution*)。其密度函数为

$$p(y;\alpha,\beta) = \binom{n}{y}\frac{B(\alpha+y, n+\beta-y)}{B(\alpha,\beta)}, y = 0,1,\cdots,n。$$

代入μ和θ,β-二项分布的密度函数为

$$p(y;\mu,\theta) = \binom{n}{y}\frac{\left[\prod_{k=0}^{y-1}(\mu+k\theta)\right]\left[\prod_{k=0}^{n-y-1}(1-\mu+k\theta)\right]}{\prod_{k=0}^{n-1}(1+k\theta)}。 \tag{13.9}$$

利用β-二项分布的矩函数可以更容易地理解该分布的特征。它的前两阶矩函数分别是

$$E(Y) = n\mu, \text{var}(Y) = n\mu(1-\mu)\left[1+(n-1)\theta/(1+\theta)\right]。$$

在β分布中,随着$\theta\to 0$,$\text{var}(\pi)\to 0$,并且该分布在μ处收敛于一个退化分布(degenerate distribution)。这时,$\text{var}(Y)\to n\mu(1-\mu)$,β-二项分布收敛于$\text{bin}(n,\mu)$。

13.3.2 β-二项分布模型

应用β-二项分布的模型允许μ(以及相应的$E(Y)$)依赖于解释变量,其中,最简单的模型是对所有观测值都令μ等于一个未知的常数(Prentice(1986)通过允许μ取决于协变量,讨论了相应的扩展)。与广义线性模型相同,在β-二项分布模型中,可以使用不同

的连结函数,其中 logit 连结最为常见。对于进行了 n_i 次试验的第 i 个观测值,假定 y_i 服从基数为 n_i 且参数为(μ_i,θ)的 β-二项分布,将 μ_i 与预测变量相连接的模型为

$$\text{Logit}(\mu_i) = \alpha + \boldsymbol{\beta}'\mathbf{x}_i。$$

即便是在 θ 已知的情况下,β-二项分布也不属于自然指数族分布。有关 β-二项分布模型的研究已经给出了多种模型拟合方法(注解 13.4)。Crowder(1978)讨论了具有 ANOVA 形式的模型的似然特性。Hinde 和 Demétrio(1998)通过迭代法来求解模型的最大似然拟合,先在给定 θ 后求解关于回归参数 $\boldsymbol{\beta}$ 的似然方程,然后在给定 $\boldsymbol{\beta}$ 后求解关于 θ 的似然方程。其中迭代的每一步都可利用 Newton-Raphson 算法。McCulloch 和 Searle (2001, p. 61)给出了独立观测值服从单一的 β-二项分布情况下,有关$(\hat{\mu},\hat{\theta})$以及$(\hat{\alpha},\hat{\beta})$的渐近协方差矩阵。

一种更为简便的方法是,利用与 β-二项分布具有相似方差函数的类似然函数来处理过度离散的二分计数。类似然方差函数可表示为

$$v(\mu_i) = n_i\mu_i(1-\mu_i)\left[1+(n_i-1)\rho\right] \tag{13.10}$$

其中 $|\rho|\leqslant 1$。尽管这个方差函数是由 β-二项模型衍生出来的,它的推导只需假定 π_i 服从 $\text{var}(\pi_i)=\rho\mu_i(1-\mu_i)$的分布。另外,它也可以通过假定相加等于 y_i 的 n_i 个二分随机变量中两两之间的相关系数 ρ 相等而推导得出。当 $\rho=0$ 时,该方差函数就简化为普通的二项分布方差函数。当 $\rho>0$ 时,就对应于过度离散的情况。

对于这种类似然法,Williams(1982)给出了估计 $\boldsymbol{\beta}$ 以及过度离散参数 ρ 的一个迭代程序。他令 $\hat{\rho}$ 的取值使得相应方差函数的皮尔逊残差平方和 X^2 等于该模型的残差自由度。这需要一个二步迭代过程:(1)给定 $\hat{\rho}$,求解关于 $\boldsymbol{\beta}$ 的类似然方程;(2)利用更新后的 $\hat{\boldsymbol{\beta}}$,在使 X^2(其值取决于 $\hat{\boldsymbol{\beta}}$ 和 $\hat{\rho}$)等于其自由度的方程中,求解 $\hat{\rho}$ 的值。

另一种类似然法使用一个更简单的方差函数

$$v(\mu_i) = \phi n_i\mu_i(1-\mu_i)。 \tag{13.11}$$

这个函数我们曾在第 4.7.3 节介绍过。当 $\phi=1.0$ 时,它等于普通的二项分布方差函数;当 $\phi>1$ 时,就对应于过度离散的情况。使用这种方法,$\hat{\boldsymbol{\beta}}$ 等于普通的二项分布模型对其进行的最大似然估计。一般来说,$\hat{\phi}=X^2/\text{df}$,其中 X^2 是二项分布模型的皮尔逊拟合统计量(Finney, 1947)。过度离散方法中的标准误等于将二项分布模型中的相应标准误乘以 $\hat{\phi}^{1/2}$。

Liang 和 McCullagh(1993)给出了分别使用这两个方差函数的几个例子。将普通二项分布模型的标准化残差与基数$\{n_i\}$进行绘图,可以用来比较哪一个函数更合适。随着 n_i 的增加,当残差的分布范围也表现出增加的趋势时,β-二项分布形式的方差函数可能更为适当。这是因为,当 β-二项分布方差函数成立时,随着 n_i 的增加,普通二项分布模型残差的分母会逐渐变得过小。当$\{n_i\}$完全相同时,这两种类似然法是等价的。只有当这些基数相差很大时,两种方法的结果才会出现明显差异。由于方差函数 $v(\mu_i)=\phi n_i\mu_i(1-\mu_i)$在 $n_i=1$ 时存在结构性问题(习题 13.33)且缺乏直接的推理依据,我们建议使用具有 β-二项分布方差函数的类似然法。

13.3.3 例子:再论畸形学中的过度离散问题

回顾表 4.5 中一项畸形学实验的结果, Liang 和 McCullagh(1993)以及 Moore 和 Tsiatis(1991)曾分析过这些数据。食用缺铁食物的雌鼠被分成了 4 组:第 1 组雌鼠只注

射安慰剂,其他组别按照不同的时间安排都注射了补铁剂。让这些雌鼠受孕,并在3周后进行解剖。结果变量为每只雌鼠所怀的每个胎儿是否已经死亡。由于存在未观察到的协变量,我们自然地允许在每个干预组中死亡概率因母鼠而异。

令 y_i 表示第 i 只母鼠所怀的 n_i 个胎儿中死亡的数量,π_{it} 表示第 i 只母鼠所怀的第 t 个胎儿的死亡概率。首先,假定 y_i 是一个 $\text{bin}(n_i, \pi_{it})$ 变量,这样有

$$\text{Logit}(\pi_{it}) = \alpha + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_4 z_{4i},$$

其中,如果第 i 只母鼠被分在第 g 组,$z_{gi}=1$,否则它等于0。该模型将属于第 g 组的母鼠所怀胎儿视为具有相同的死亡概率,即都等于 $\exp(\alpha+\beta_g)/[1+\exp(\alpha+\beta_g)]$,其中 $\beta_1 = 0$。然而,这里存在过度离散的问题,模型对应的 $X^2 = 154.7$, $G^2 = 173.5(\text{df}=54)$。表13.5给出了该模型的最大似然估计及其标准误。

表 13.5 通过多个 Logit 模型对表 4.5 数据进行拟合的估计结果

参数	Logit 模型的类型[a]				
	二项分布最大似然估计	类似然估计(1)	类似然估计(2)	GEE	广义线性混合模型
截距	1.144(0.129)	1.212(0.223)	1.144(0.219)	1.144(0.276)	1.802(0.362)
第2组	-3.322(0.331)	-3.370(0.563)	-3.322(0.560)	-3.322(0.440)	-4.515(0.736)
第3组	-4.476(0.731)	-4.585(1.303)	-4.476(1.238)	-4.476(0.610)	-5.855(1.190)
第4组	-4.130(0.476)	-4.250(0.848)	-4.130(0.806)	-4.130(0.576)	-5.594(0.919)
过度离散	无	$\hat{\rho}=0.192$	$\hat{\phi}=2.86$	$\hat{\rho}=0.185$	$\hat{\sigma}=1.53$

a 二项分布最大似然估计假定不存在过度离散,类似然估计(1)是基于有 β-二项分布方差函数的类似然法,类似然估计(2)是基于对二项分布方差函数进行膨胀的类似然法;类似然估计(2)和 GEE(独立的操作性相关结构)估计值与二项分布最大似然估计相同。括号中的数值为标准误。

表13.5也给出了运用两种类似然方法拟合的结果。二者所得到的参数估计值和标准误基本相似。对于使用方差函数 $v(\mu_i) = \phi n_i \mu_i(1-\mu_i)$ 的类似然法,其估计值与二项分布最大似然估计值完全相同;但是它的标准误等于二项分布模型相应标准误的 $\hat{\phi}^{1/2} = (X^2/\text{df})^{1/2} = \sqrt{154.7/54} = 1.69$ 倍。在使用 β-二项分布方差函数的类似然估计中,$\hat{\rho} = 0.192$。这个拟合过程假定 Y_i 的方差等于

$$n_i \mu_i (1-\mu_i)\left[1 + 0.192(n_i - 1)\right]。$$

当 $n_i = 6$ 时,它相当于二项分布方差的两倍;当 $n_i = 11$ 时,它大约是二项分布方差的三倍。表13.5的结果表明,即便对过度离散进行了如上调整,每个干预组的死亡概率仍然显著低于注射安慰剂的控制组。

图13.4将二项分布 Logit 模型的标准化皮尔逊残差与每窝所包含的胎儿数量进行了绘图。随着每窝所含胎儿数量的增加,残差的变动性也明显增大,这表明在此使用 β-二项分布方差函数是合理的。该方差函数中的 ρ 对应于 β-二项分布方差中的 $\theta/(1+\theta)$。在该分布中,或者更一般地说,$\hat{\rho}=0.192$ 意味着,对同属于某一组的母鼠的胎儿而言,所估计的死亡概率的标准差为 $\sqrt{0.192\mu_i(1-\mu_i)}$。当均值为0.5时,该标准差等于0.22;当均值为0.1或0.9时,它等于0.13。这一结果反映了很强的异质性。更一般的情况下,可以令模型中各干预组具有不同的 ρ,或者控制组与干预组的 ρ 可以取不同值。我们将这个问题留给读者自行完成。

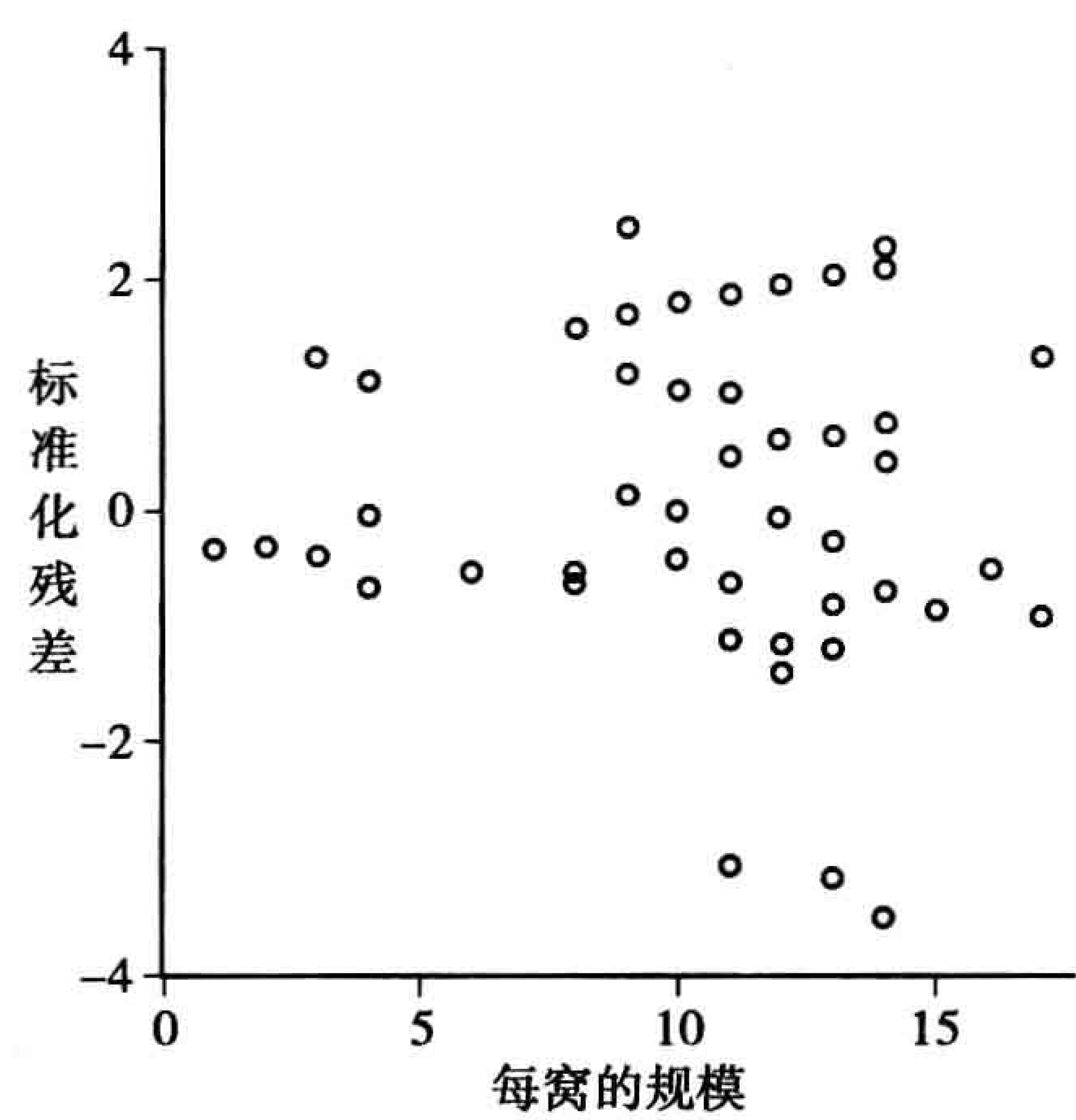


图 13.4 对表 4.5 数据拟合的二项分布 Logit 模型的标准化皮尔逊残差

为了便于比较,表 13.5 还给出了使用 GEE 法拟合这个 Logit 模型的结果,在拟合过程中假定同一窝内的观测值之间具有独立的操作性相关结构。其估计值与二项分布 Logit 模型的最大似然估计相同,但该方法的经验调整给出的标准误较大。使用可交换的操作性相关结构得到的结果与此相似,且后者估计的关于二分结果变量的窝内相关系数为 0.185。这与具有 β-二项分布方差函数的类似然法所给出的 0.192 相当。GEE 法中的标准误与通过类似然法得到的标准误有所不同。这可能是因为样本规模不足以支持 GEE 的 sandwich 调整,这种调整一般会低估标准误,除非群组的数量很大。或者,这种差别也可能只是反映 GEE 使用的方差函数与类似然法不同。

最后,表 13.5 也给出了在二项分布 Logit 模型中加入关于第 i 窝的正态随机截距 u_i 的广义线性混合模型的结果。就干预组与控制组的相对显著性水平而言,模型结果与前面的分析基本相似。这个 logistic-正态模型所估计的效应更大一些,因为这些效应是对象别的(即分窝别的),而不是总体平均效应。

13.3.4 共轭混合模型

β-二项分布模型是*共轭混合模型*(*conjugate mixture model*)的一个例子。这些模型具有封闭形式的边际分布,以某个参数为条件,数据服从特定的分布,此外,参数本身的分布具有封闭形式的边际分布。

与之相类似,在贝叶斯方法中,共轭先验分布就是一种将先验分布与似然函数相结合,由此推出一个封闭形式的后验分布的分布。例如,对于参数服从 β 先验分布的二项分布观测值来说,其参数的后验分布仍然服从 β 分布。在一些专门用于求解决定后验分布的复杂积分的运算方法(如马尔科夫蒙特卡洛法)得到运用之前,共轭模型是进行贝叶斯分析的主要方法。

β-二项分布共轭混合模型以二分结果变量的试验总数为基础进行分析。在下一节中,我们将介绍关于计数数据的共轭混合模型。该模型设定混合的泊松分布参数服从 γ 分布。共轭混合方法的一个缺点在于,它缺乏一般性和灵活性,每一类型的问题都需要不同的混合分布。另外,模型所允许的额外变动性不一定按照与普通预测变量相同的方式进入模型,并且在共轭混合模型中多元随机效应的结构难以设定。Lee 和 Nelder (1996)对此方法进行了讨论,并提到了随机效应不服从正态分布的多种广义线性混合模型。

13.4 负二项回归

负二项(negative binomial)分布是关于计数数据的一个共轭混合分布。当泊松广义线性模型出现过度离散的问题时,可以考虑使用负二项模型。

13.4.1 作为泊松分布的 γ 混合的负二项分布

在第 4.3.3 节我们提到,泊松模型的一个严重缺陷在于 Y 的方差必须等于其均值。因而,对于给定的均值,方差不能随着更多的预测变量加入模型而下降。计数数据经常出现过度离散的情况,即它的方差大于均值。例如,当某些相关联的解释变量未被包括进模型中时,就可能会出现这种情况。混合模型是处理过度离散问题的一种灵活方式,在模型中预测变量的某个取值组合处,给定均值后 Y 服从泊松分布,但允许均值本身按照某个分布变动。

假定(1)给定 λ,Y 服从均值为 λ 的泊松分布;并且(2)λ 服从 γ 分布 $G(k,\mu)$。关于 λ 的 γ 概率密度函数为

$$f(\lambda;k,\mu) = \frac{(k/\mu)^k}{\Gamma(k)}\exp(-k\lambda/\mu)\lambda^{k-1}, \lambda \geqslant 0。\qquad (13.12)$$

这个 γ 分布具有

$$E(\lambda) = \mu, \mathrm{var}(\lambda) = \mu^2/k。$$

参数 $k>0$ 描述了函数的形状。该密度函数向右偏斜,其偏斜度随着 k 的上升而下降。

在边际上,泊松分布的 γ 混合对应于 Y 的负二项分布,其概率密度函数为

$$p(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y, y = 0,1,2,\cdots。\qquad (13.13)$$

这个负二项分布具有

$$E(Y) = \mu, \mathrm{var}(Y) = \mu + \mu^2/k。$$

在这里,k^{-1}被称为离散参数(*dispersion parameter*)。随着 $k^{-1}\to 0$,γ 分布具有 $\mathrm{var}(\lambda)\to 0$,并收敛于一个在 μ 处的退化分布;相应地,负二项分布具有 $\mathrm{var}(Y)\to\mu$,并收敛于均值为 μ 的泊松分布。

对于给定的 k^{-1},负二项分布属于自然指数族,它的自然参数为 $\log[\mu/(\mu+k)]$。然而,通常情况下,离散参数 k^{-1}本身是未知的,对它进行估计有助于了解过度离散的程度。k^{-1}越大,与普通泊松广义线性模型相比过度离散的问题就越严重。在观测值相互独立的情况下,关于 μ 的最大似然估计为样本均值,但对 k^{-1}的最大似然估计需要使用迭代法(在 1953 年 C. Bliss 发表于 *Biometrics* 的一篇文章的附录中,R. A. Fisher 给出了证明)。习题 13.40 给出了关于 γ 的另一种参数化方法,它意味着负二项分布具有一个线性的方差函数,而不是二次项函数。

13.4.2 负二项回归模型

关于计数数据的负二项模型允许 μ 的值取决于解释变量(Lawless, 1987)。一般情况下,在这类模型中 k^{-1}的取值对所有的观测值都相同。这样,γ 混合分布的变动系数(coefficient of variation)等于一个常数,$\sqrt{\mathrm{var}(\lambda)}/\mathrm{E}(\lambda) = 1/\sqrt{k}$,即其标准差随着均值的增加而增加。与泊松对数线性模型一样,负二项模型中最常用的连结函数是对数连结。有时,也可以考虑使用恒等连结,比如当模型只包括一个分类预测变量的时候。

在 k 固定的情况下，负二项模型是一个广义线性模型。因此，关于回归参数 $\boldsymbol{\beta}$ 的似然方程是方差函数为 $v(\mu)=\mu+\mu^2/k$ 的普通广义线性模型的似然方程（参见式 4.22）的一个特例。利用通常的迭代再加权最小二乘法，可以求解该模型的最大似然拟合。当 k 未知时，可以使用 Newton-Raphson 算法对所有参数同时进行最大似然拟合。或者，也可以求解当 k 取不同值时的剖面似然函数（Lawless，1987）。还有一种方法是在以下两步之间进行迭代转换，直到收敛为止：(1) 对于给定的 k，利用迭代再加权最小二乘法求解关于 $\boldsymbol{\beta}$ 的方程；(2) 对于给定的 $\hat{\boldsymbol{\beta}}$，利用 Newton-Raphson 算法来估计 k。

负二项模型的整个对数似然函数 $L(\boldsymbol{\beta},k;\mathbf{y})$ 满足

$$\frac{\partial^2 L}{\partial\beta_j\partial k}=\sum_i \frac{y_i-\mu_i}{(k+\mu_i)^2 g'(\mu_i)}x_{it}。$$

因此，对于每个 j，都有 $E(\partial^2 L/\partial\beta_j\partial k)=0$。相似地，在期望信息矩阵的逆矩阵中，连接 k 与每个 β_j 的元素均为 0；由于它是一个渐近协方差矩阵，$\hat{\boldsymbol{\beta}}$ 与 $\hat{k}$ 渐近独立。由此推出，从上述迭代过程的第（1）部分所得到的关于 $\hat{\boldsymbol{\beta}}$ 的标准误是正确的。Cameron 和 Trivedi（1998, p. 72）推导了这个渐近协方差矩阵，并（另见 Lawless（1987））考虑了关于 k^{-1} 的矩估计，以及估计结果的稳健性。他们指出，如果关于均值的模型设定是正确的，即便真实的分布不是负二项分布，这个模型的 $\hat{\boldsymbol{\beta}}$ 仍然具有一致性。

13.4.3 例子：认识凶杀受害者的频数

表 13.6 给出了 1 308 名调查对象关于以下问题的回答结果：在过去的 12 个月中，你所认识的人里面有多少遭遇凶杀身亡？该表按照被访者自报的种族（白人或者黑人）将回答结果进行了划分。其中，159 名黑人被访者所报告的样本均值为 0.522，方差为 1.150。1 149 名白人的样本均值为 0.092，方差为 0.155。

表 13.6 按种族划分的认识的人中过去一年遭遇凶杀人数，以及泊松模型与负二项模型的拟合结果

	数据		泊松广义线性模型		负二项广义线性模型		泊松广义线性混合模型	
回答结果	黑人	白人	黑人	白人	黑人	白人	黑人	白人
0	119	1 070	94.3	1 047.7	122.8	1 064.9	116.7	1 068.3
1	16	60	49.2	96.7	17.9	67.5	24.5	65.3
2	12	14	12.9	4.5	7.8	12.7	8.1	10.1
3	7	4	2.2	0.1	4.1	2.9	3.6	2.8
4	3	0	0.3	0.0	2.4	0.7	1.9	1.1
5	2	0	0.0	0.0	1.4	0.2	1.1	0.5
6	0	1	0.0	0.0	0.9	0.1	0.7	0.3

来源：1990 年综合社会调查，美国民意研究中心。

对计数数据进行模型分析，首选的模型自然是泊松广义线性模型，如以种族的虚拟变量作为预测变量的对数线性模型。令 y_{it} 表示属于种族 i 的第 t 个对象的回答。对于 $\mu_{it}=E(Y_{it})$，这个模型可表示为

$$\log\mu_{it}=\alpha+\beta x_{it},$$

其中，$x_{1t}=1$（黑人），$x_{2t}=0$（白人）。该模型的拟合结果为 $\log\hat{\mu}_{it}=-2.38+1.733x_{it}$。它

所估计的期望结果对黑人而言等于 exp(−2. 38 + 1. 733) = 0. 522,对白人而言等于 exp(−2. 38) = 0. 092,即各自的样本均值。在这个模型中,无论使用哪种连结函数,似然方程决定了所拟合的均值一定等于样本均值。由于 $\hat{\beta}=1.733$(SE = 0. 147),即为黑人与白人之间的对数均值之差,两个样本均值的比率为 exp(1. 733) = 5. 7 = 0. 522/0. 092。然而,对于每个种族,样本方差大约是其均值的两倍。表 13. 6 给出了这个模型的拟合情况。与泊松广义线性模型对应的预测值相比,在 $y=0$ 以及 y 取较大值时的观测计数更大,这表明存在过度离散的问题。

另一种可选的分析方法是保持模型形式不变,但假定结果变量服从负二项分布。在这里,使用混合模型是合理的,由于各种人口学因素的影响,不同种族的调查对象对应的 Y 的分布可能确实存在异质性。最大似然拟合结果表明,与普通泊松广义线性模型($k^{-1}=0$时的特例)相比,负二项模型的偏离度下降了 122. 2。表 13. 6 也给出了这个模型的拟合结果,它在 $y=0$ 和 1 处的预测结果有了显著改善。

表 13. 7 给出了负二项模型和泊松广义线性模型的参数估计。两个模型都有 $\hat{\beta}=1.733$,这是因为它们拟合的均值都等于样本均值。但是,关于 $\hat{\beta}$ 的标准误的估计值在泊松广义线性模型中为 0. 147,在负二项模型中上升到 0. 238。相应地,有关黑人与白人之间的均值比率的 95% 的沃尔德置信区间在泊松广义线性模型中为 exp[1. 733 ± 1. 96(0. 147)] = (4. 2,7. 5),在负二项模型中增大到 exp[1. 733 ± 1. 96(0. 238)] = (3. 5, 9. 0)。这些结果在考虑了过度离散的问题后,便不再像起初的模型所显示得那么精确。

在这个负二项模型中,$\hat{k}^{-1}=4.94$(SE = 1. 00)。存在强有力的证据表明 $k^{-1}>0$,因而,负二项模型比泊松广义线性模型更合适。它所估计的 Y 的方差为 $\hat{\mu}+\hat{\mu}^2/\hat{k}=\hat{\mu}+4.94\,\hat{\mu}^2$,也即,对白人等于 0. 13,对黑人等于 1. 87。与泊松模型相比,这个结果明显更接近于样本值。

表 13. 7 还给出了在使用恒等连结的情况下负二项模型和泊松模型的拟合结果。拟合方程 $\hat{\mu}_{it}=0.092+0.430x_{it}$ 复制了样本均值。在这里,$\hat{\beta}$ 指的是均值之差,而不是均值的对数之比。模型所估计的均值之差 $\hat{\beta}=0.430$,在泊松模型中其标准误为 SE = 0. 058,而在负二项模型中为 SE = 0. 109。相对来说,负二项模型的结果更不精确,但也更为现实。在恒等连结的情况下,负二项模型所估计的离散参数仍为 $\hat{k}^{-1}=4.94$。

表 13. 7　对凶杀数据所拟合的模型的参数估计

项	对数连结的模型			恒等连结的模型	
	负二项广义线性模型	泊松广义线性模型	泊松广义线性混合模型	负二项广义线性模型	泊松广义线性模型
α	−2. 38	−2. 38	−3. 69	0. 092	0. 092
β	1. 733	1. 733	1. 897	0. 430	0. 430
SE($\hat{\beta}$)	0. 238	0. 147	0. 246	0. 109	0. 058

13.5　包括随机效应的泊松回归

在第 12 章中,我们介绍了关于分类结果变量的广义线性混合模型。对于其他类型

的离散结果变量，如计数变量，也可以应用广义线性混合模型。在本节中，我们介绍关于计数数据的泊松回归模型。

我们已经看到，混合模型是一种处理过度离散问题的灵活方式。在第 13.4 节中，我们利用 γ 分布对泊松分布进行了混合，从而在边际上得出了负二项模型。Breslow(1984)以及 Hinde(1982)建议，可以考虑具有对数连结和正态随机截距的广义线性混合模型(式 12.1)。对于在第 i 个群组中的第 t 个观测值的均值，该模型可表示为

$$\log\left[E(Y_{it} \mid u_i)\right] = \mathbf{x}'_{it}\boldsymbol{\beta} + u_i, \tag{13.14}$$

其中，$\{u_i\}$ 服从独立的 $N(0,\sigma^2)$ 分布。以 u_i 为条件，y_{it} 服从泊松分布。只要 $\sigma>0$，该泊松分布的方差就在边际上大于其均值。

有关泊松广义线性混合模型的应用包括，流行病学中对癌症发病率地图的分析(Breslow and Clayton, 1993)以及对细菌数量变动性的模型分析(Aitchison and Ho, 1989)。尽管有可能使用除对数连结以外的其他连结函数，恒等连结(以及所有取值范围只能为正实数的其他连结函数)存在一个结构性问题，即在对应的正态随机效应 $\sigma>0$ 的情况下，线性预测项有取负值的可能性，但是泊松变量的均值必须是非负的。

负二项模型(对于给定的 k)是一个具有非正态随机效应的广义线性混合模型。在对数连结的情况下，负二项模型相当于一个形式为式 13.14 的，包括随机截距的对数线性模型，其中 $\exp(u_i)$ 服从均值为 1 和方差为 k^{-1} 的 γ 分布。在恒等连结的情况下，负二项模型通常比泊松广义线性混合模型更好。无论取什么样的 γ 混合分布，负二项模型中的相应边际均值总是非负的。

13.5.1　泊松广义线性混合模型所隐含的边际模型

通过求平均消除随机效应，泊松广义线性混合模型(式 13.14)隐含着一个相对简单的边际模型。边际分布的均值为

$$E(Y_{it}) = E\left[E(Y_{it} \mid u_i)\right] = E\left[e^{\mathbf{x}'_{it}\boldsymbol{\beta}+u_i}\right] = e^{\mathbf{x}'_{it}\boldsymbol{\beta}+\sigma^2/2}。$$

这里，因为服从 $N(0,\sigma^2)$ 分布的变量 u_i 的矩量生成函数为 $E[\exp(tu_i)]=\exp(t^2\sigma^2/2)$，所以 $E[\exp(u_i)]=\exp(\sigma^2/2)$。因此，对于泊松广义线性混合模型，条件均值的对数等于 $\mathbf{x}'_{it}\boldsymbol{\beta}+u_i$，而边际均值的对数等于 $\mathbf{x}'_{it}\boldsymbol{\beta}+\sigma^2/2$。这时，对数线性模型仍然适用，解释变量的边际效应与群组别(cluster-specific)效应相同。由此得出，在条件模型与边际模型中，在 $\mathbf{x}'_{it}$ 的两个不同取值处的均值比率都相等。不过，在边际上，截距被抵消(注意：由于连结函数是非线性函数，可以应用詹生不等式(Jensen's inequality))。

边际分布的方差为

$$\begin{aligned}\operatorname{var}(Y_{it}) &= E\left[\operatorname{var}(Y_{it} \mid u_i)\right] + \operatorname{var}\left[E(Y_{it} \mid u_i)\right] = E\left[e^{\mathbf{x}'_{it}\mathbf{B}+u_i}\right] + e^{2\mathbf{x}'_{it}\boldsymbol{\beta}}\operatorname{var}(e^{u_i})\\ &= e^{\mathbf{x}'_{it}\boldsymbol{\beta}+\sigma^2/2} + e^{2\mathbf{x}'_{it}\boldsymbol{\beta}}(e^{2\sigma^2}-e^{\sigma^2}) = E(Y_{it}) + \left[E(Y_{it})\right]^2(e^{\sigma^2}-1)。\end{aligned}$$

在这里，通过求解 $t=2$ 和 $t=1$ 时的矩量生成函数可得，$\operatorname{var}(e^{u_i})=E(e^{2u_i})-[E(e^{u_i})]^2=e^{2\sigma^2}-e^{\sigma^2}$。与负二项模型相同，边际方差是边际均值的二次函数。当 $\sigma>0$ 时，它大于边际均值；当 $\sigma=0$ 时，就得到了普通泊松模型。而当 $\sigma>0$ 时，边际分布并不服从泊松分布，且随着 σ 的上升，方差超出均值的程度也上升。

如同二分的广义线性混合模型，给定 u_i 后，Y_{it} 和 Y_{is} 相互独立，但二者在边际上存在非负相关关系。对于 $t\neq s$，

$$\text{cov}(Y_{it},Y_{is}) = E\left[\text{cov}(Y_{it},Y_{is} \mid u_i)\right] + \text{cov}\left[E(Y_{it} \mid u_i),E(Y_{is} \mid u_i)\right]$$
$$= 0 + \text{cov}\left[\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i),\exp(\mathbf{x}'_{is}\boldsymbol{\beta} + u_i)\right]。 \quad (13.15)$$

以上协方差中最后一项的函数都是关于 u_i 的单调增函数,因而它们之间具有非负相关关系(习题 13.44)。

13.5.2 例子:所认识的凶杀受害者的频数

现在我们回到表 13.6 的数据,即按种族划分的所认识的人在过去 12 个月中遭遇凶杀的人数。允许对象间异质性的模型是合理的选择。对于属于种族 i 的第 t 个对象,泊松广义线性混合模型为

$$\log\left[E(Y_{it} \mid u_{it})\right] = \alpha + \beta x_{it} + u_{it},$$

其中$\{u_{it}\}$服从独立的 $N(0,\sigma^2)$分布。对于白人,对数均值按照 $N(\alpha,\sigma^2)$分布变动;对于黑人,对数均值按照 $N(\alpha+\beta,\sigma^2)$变动。给定 u_{it},y_{it}服从泊松分布。

表 13.6 也给出了这个模型的拟合情况,表 13.7 给出了相应的参数估计,其中,随机效应具有 $\hat{\sigma}=1.63$(SE = 0.15)。与普通泊松广义线性模型相比,它的偏离度下降了 116.6,这表明允许异质性的模型对数据拟合得更好。就随机效应分布的均值($u_{it}=0$)对应的研究对象来说,对黑人估计的期望值等于 exp(−3.69 + 1.90) = 0.167,对白人等于 exp(−3.69) = 0.025。模型所拟合的黑人的边际均值等于 $\exp(\hat{\sigma}+\hat{\beta}x_{it}+\hat{\sigma}^2/2)$即 0.63,白人的边际均值等于 0.09。黑人的边际方差为 0.21,而白人为 5.78。这些值比样本均值和方差都要大一些,可能的原因是,所拟合的分布在最大观察值 6 之上的取值可能性仍不可忽略。

13.5.3 负二项模型与泊松广义线性混合模型的比较

与负二项模型相比,包括正态随机效应的泊松广义线性混合模型具有易于处理多元随机效应以及多层次模型的优点。但是,相对而言,负二项模型的解释更加容易。我们已经看到,在负二项模型中可以使用恒等连结,这对于一些简单的例子——如前面只包括一个分类预测变量的情况,非常有意义。无论使用哪种连结函数,在只有一个分类预测变量的情况下,负二项模型的拟合均值一定等于样本均值。而对于泊松广义线性混合模型来说,这一点并不成立。

除了泊松广义线性混合模型和负二项模型外,另一种处理计数数据的过度离散问题的方法是使用具有方差函数

$$v(\mu_i) = \phi\mu_i$$

的类似然法,其中 ϕ 是一个常数。一般来说,在探索性分析中,这种方法就足够了。

注 解

第 13.1 节:潜类模型

13.1 讨论有关潜类以及相应的潜变量模型的拟合与解释参见:Aitkin et al.(1981)、Bartholomew and Knott(1999)、Clogg(1995)、Clogg and Goodman(1984)、Goodman(1974)、Haberman(1979, Chap. 10)、Hagenaars(1998)、Heinen(1996)、Lazarfeld and Henry(1968)。

13.2 Rudas 等(1994)提出了一种更为聪明的混合方法来描述拟合优度。对于真实概率

为 $\boldsymbol{\pi}$ 的列联表拟合的模型 M,他们使用混合概率 $\boldsymbol{\pi}=(1-\rho)\boldsymbol{\pi}_1+\rho\boldsymbol{\pi}_2$,其中 $\boldsymbol{\pi}_1$ 是基于模型的概率,而 $\boldsymbol{\pi}_2$ 是未加限定的概率。他们关于拟合不足的指标是满足上式的最小可能的 ρ 值,它表示总体中无法由该模型所描述的比例。这反映了任何一个模型都不可能真正成立,但只要 ρ 接近于0,模型仍是很有意义的。这种混合方法与潜类模型不同,后者中 $\boldsymbol{\pi}_1$ 和 $\boldsymbol{\pi}_2$ 都对应着独立性的情况。

第 13.2 节:非参数随机效应模型

13.3 有关 Rasch 模型与准对称性模型之间的联系,参见:Agresti(1993)、Conaway(1989)、Darroch(1981)、Darroch et al.(1993)、Hatzinger(1989)、Kelderman(1984)。在关于配对数据的随机效应模型(式 12.16)中,对(u_{i1},u_{i2})的非参数或条件最大似然估计对应于一个多元准对称性模型(Agresti,1997)。正态随机效应之间存在相关关系的模型(式 12.16)是由 Goodman(1974)提出的离散潜类模型的连续形式,该离散模型基于两个相关的二分潜变量。

第 13.3 节:β-二项分布模型

13.4 Skellam(1948)介绍了 β-二项分布,并讨论了有关的参数估计。对于使用这个分布或相应的类似然法的模型分析,参见:Brooks et al.(1997)、Crowder(1978)、Hinde(1996)、Lee and Nelder(1996)、Liang and Hanfelt(1994)、Liang and McCullagh(1993)、Lindsey and Altham(1998)、Moore(1986a)、Moore and Tsiatis(1991)、Nelder and Pregibon(1987)、Prentice(1986)、Rosner(1984,1989)(包括 Neuhaus and Jewell(1990a)的评论)、Slaton et al.(2000)、Williams(1975,1982)。有关使用 β-二项分布方差函数的类似然法,Ryan(1995)以及 Williams(1988)给出了它与最大似然法相比所具有的优点。通常来说,允许类似然函数的规模参数 ρ(或者 β-二项分布中的相应参数 θ)在不同组之间变动是有意义的。

β-二项分布可以扩展为 Dirichlet-多项分布。以概率为条件,该分布是一个多项分布。而概率本身则服从 Dirichlet 分布,它是对所定义的相加为 1 的概率向量 $\boldsymbol{\beta}$ 的一个扩展。参见:Mosimann(1962)、Paul et al.(1989)。

13.5 Kupper 等(1986)以及 Ryan(1992)讨论了在发育毒性研究中由于窝群效应所引起的过度离散问题以及相应的模型分析。有关内容,可参见:Follman and Lambert(1989)、Kupper and Haseman(1978)、Lefkopulou et al.(1989)。

第 13.4 节:负二项回归

13.6 Greenwood 和 Yule(1920)通过关于泊松分布的 γ 混合推导出了负二项分布。Johnson 等(1992)总结了该分布的特性;讨论利用该分布所进行的模型分析可参见:Biggeri(1998)、Cameron and Trivedi(1998)、Hinde and Demétrio(1998)、Lawless(1987)。

习题

应用部分

13.1 对于有关合法流产态度的 2^3 表格(表 10.13),合并分性别的数据,并拟合一个具有两个潜类的潜类模型。证明这个模型是饱和模型。对其中的每个潜类,报告所估计的在三种情形下分别支持合法流产的概率。尝试解释每个潜类的意义。

13.2 利用一个 $q=2$ 的潜类模型分析表 8.3。

a. 对于第一个潜类的对象,估计出现以下情况的概率:(**i**)服用大麻,(**ii**)饮酒,(**iii**)吸烟,(**iv**)三种行为都有,(**v**)三种行为都没有。

b. 给定一个对象具有以下情况:(**i**)服用大麻,(**ii**)饮酒,(**iii**)吸烟,(**iv**)三种行为

都有,(v)三种行为都没有,估计他属于第一个潜类的概率。

13.3 通过潜类模型分析表 8.19 中有关政府花费的数据。

13.4 对于捕获-再捕获实验,Coull 和 Agresti(1999)使用了一个具有可交换关联且不包括高阶项的对数线性模型,说明为什么模型的期望频数满足

$$\log \mu(y_1,\cdots,y_T) = \lambda + \beta_1 y_1 + \cdots + \beta_T y_T + \beta(y_1 y_2 + y_1 y_3 + \cdots + y_{T-1} y_T)。$$

证明该模型对表 12.6 的拟合结果为 $\hat{N}=90.5$,以及关于 N 的 95% 的剖面似然置信区间为(75,125)。

13.5 利用软件或编写程序来复制第 13.2 节中关于合法流产态度的分析:(a)拟合非参数随机效应 Logit 模型(式 13.3),(b)准对称性模型。

13.6 在某一关于佛罗里达州中北部 13 县的 18 岁以下女孩怀孕率的数据中,包括每个县 18 岁以下女孩的 n_i = 生育数量以及 y_i = 总人数,数字均为三年之和(参见:J. Booth, in *Statistical Modelling: Lecture Notes in Statistics*, 104, Springer, 43-52, 1995)。

a. β-二项分布模型表示,给定$\{\pi_i\}$后,$\{Y_i\}$是独立的$\{\text{bin}(n_i,\pi_i)\}$变量,并且$\{\pi_i\}$服从独立的$\beta(\alpha,\beta)$分布。相应的参数最大似然估计结果为 $\hat{\alpha}=9.9$ 和 $\hat{\beta}=240.8$(感谢 J. Booth 所做的分析)。利用均值和方差来描述所估计的 β 分布以及 Y_i 的边际分布(表示为 n_i 的函数)。

b. 利用具有方差函数(式 13.10)的类似然法拟合模型 $\text{Logit}(\mu_i)=\alpha$,结果为 $\hat{\alpha}=-3.18$和 $\hat{\rho}=0.005$。给出关于 Y_i 的均值和方差的估计值。

c. 利用具有方差函数(式 13.11)的类似然法拟合模型 $\text{Logit}(\mu_i)=\alpha$,结果为 $\hat{\alpha}=-3.35$和 $\hat{\phi}=8.3$。给出关于 Y_i 的均值和方差的估计值。

d. Logistic-正态广义线性混合模型 $\text{Logit}(\pi_i)=\alpha+u_i$ 的拟合结果为 $\hat{\alpha}=-3.24$ 和 $\hat{\sigma}=0.33$。给出关于 Y_i 的均值的估计值(回顾式 12.8)。

13.7 在习题 12.2 中关于沙克·奥尼尔罚篮情况的数据,简单的二项分布模型 $\pi_i=\alpha$ 对数据拟合得不充分。拟合一个 β-二项分布模型或者使用具有该方差结构的类似然法分析此数据。利用拟合结果描述他的罚篮情况,并给出 π_i 的均值和标准差的估计值。

13.8 对于表 12.9 的毒性研究数据,将结果变量合并为一个二分变量,对胎儿状态为正常的概率构建线性 Logit 模型。

a. 普通的二项分布模型有没有显示出存在过度离散的问题?

b. 利用类似然法拟合线性 Logit 模型,允许二项分布方差膨大。标准误发生了怎样的变化?

c. 利用类似然法拟合线性 Logit 模型,应用 β-二项分布方差函数。解释结果,并与前面的模型进行对比。

d. 利用 GEE 法拟合线性 Logit 模型,设定同一窝中的胎儿间具有可交换的操作性相关结构。解释结果,并与前面的模型进行对比,包括 GEE 法估计的相关系数和(c)部分的 $\hat{\rho}$ 的比较。

e. 在加入窝别随机效应后,拟合线性 Logit 模型。解释结果,并与前面的模型进行对比。

13.9 将第 13.3 节中有关畸形学数据(表 4.5)的各种分析进行下列扩展:

a. 加入每窝所含胎儿数量（以及组别）的预测变量。解释结果，并与不包括该变量的结果进行对比。

b. 拟合一个具有β-二项分布方差的模型（式 13.10），其中，ρ 可以在各干预组之间取不同的值。利用有关结果，推导一个只允许控制组中出现过度离散的模型。解释结果，并与在每组中 ρ 都取同一值的模型结果进行对比。

13.10　表 13.8 给出了一项在三种不同环境下鱼卵孵化结果的研究。七袋鱼卵被随机分配给三个实验组，结果变量为到第 10 天时鱼卵是否已经孵化出来。三种实验环境分别为：(1) 去除二氧化碳和氧气，(2) 只去除二氧化碳，(3) 都不去除。

表 13.8　习题 13.10 的数据

	实验组 1		实验组 2		实验组 3	
鱼卵袋	孵化数	总　数	孵化数	总　数	孵化数	总　数
1	0	6	3	6	0	6
2	0	13	0	13	0	13
3	0	10	8	10	6	9
4	0	16	10	16	9	16
5	0	32	25	28	23	30
6	0	7	7	7	5	7
7	0	21	10	20	4	20

来源：数据由佛罗里达大学动物学系的 Becca Hale 友情提供。

a. 令 π_{it}表示在第 t 个实验组中第 i 袋的一个鱼卵被孵化的概率。假定这些观测值服从独立的二项分布，拟合模型

$$\text{Logit}(\pi_{it}) = \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3,$$

其中，对于第 t 个实验组 $z_t = 1$，否则为 0。你使用的软件给出的 $\hat{\beta}_1$ 等于多少？它应当等于多少？（*提示：*注意第 1 个实验组不存在成功的情况）。

b. 利用一种允许过度离散的方法分析这些数据。解释结果，并对第 2 个和第 3 个实验组，指出是否出现了过度离散的情况。

13.11　对于习题 9.19 的火车事故例子，负二项模型假定在 14 年间事故的对数发生率都保持不变，它的估计结果为 -4.177（SE = 0.153），且估计的离散参数为 0.012。对这些结果加以解释。

13.12　在 1990 年综合社会调查（General Social Survey）中，其中一个问题要求被访者回答上个月发生性行为的次数。表 13.9 给出了按性别划分的回答结果。

表 13.9　习题 13.12 的数据

回答结果	男性	女性	回答结果	男性	女性	回答结果	男性	女性
0	65	128	9	2	2	20	7	6
1	11	17	10	24	13	22	0	1
2	13	23	12	6	10	23	0	1
3	14	16	13	3	3	24	1	0
4	26	19	14	0	1	25	1	3
5	13	17	15	3	10	27	0	1
6	15	17	16	3	1	30	3	1
7	7	3	17	0	1	50	1	0
8	21	15	18	0	1	60	1	0

来源：1990 年综合社会调查，美国民意研究中心。

a. 男性的样本均值为5.9,而女性为4.3,相应的样本方差分别为54.8和34.4。每个性别回答的众数都等于0。这里使用普通的泊松广义线性模型合适吗?加以说明。

b. 使用具有对数连结的泊松广义线性模型,以性别的虚拟变量(1 = 男性,0 = 女性)为解释变量,它估计的性别效应为0.308(SE = 0.038)。说明为什么这意味着所拟合的均值之间的比率为1.36(这也是样本均值之间的比率,因为该模型的拟合均值等于样本均值)。证明关于男性和女性均值比率的95%的沃尔德置信区间为(1.26,1.47)。

c. 与普通泊松模型相比,负二项模型的对数似然值上升了248.7(偏离度下降了497.3)。负二项模型所估计的对数均值之差也是0.308,但在这里SE = 0.127。证明关于均值比率的95%的置信区间为(1.06,1.75)。将其与泊松广义线性模型的结果进行对比,并加以解释。

d. 泊松分布的众数是其均值的整数部分,而不是0。因而,更现实的混合模型或许应当假定对于性别 i,有 ρ_i 的比例服从均值为0的泊松分布,$1-\rho_i$ 的比例服从泊松分布的 γ 混合,给出理由。说明为什么每个性别的相应边际分布是一个在0点处的退化分布和一个负二项分布的混合分布。

13.13 参见习题13.12。利用恒等连结拟合泊松和负二项广义线性模型。展示在两个广义线性模型中,所估计的男女均值之差都相等,但它们的标准误差别很大。解释原因。选择更适当的模型来构建一个关于均值之差的置信区间。

13.14 对于表4.3中马蹄蟹同伴的计数数据,表13.10给出了相应的负二项模型的最大似然拟合结果,其中以壳宽作为预测变量并使用恒等连结。

表13.10 习题13.14的数据

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
Intercept	−11.147 1	2.827 5	−16.689 0	−5.605 2	15.54
width	0.530 8	0.113 2	0.308 9	0.752 8	21.97
Dispersion	0.984 3	0.182 2	0.684 7	1.414 9	

译者注——Width:壳宽;Dispersion:离散参数。

a. 写出预测方程,并加以解释。

b. 对于预测值 $\hat{\mu}$,证明所估计的方差大约等于 $\hat{\mu}+\hat{\mu}^2$。

c. 相应的泊松广义线性模型的拟合结果为 $\hat{\mu} = -11.53 + 0.55x$(SE = 0.06)。比较两个模型中关于斜率的95%的置信区间。加以解释,并指出对于泊松广义线性模型而言,是否出现了过度离散的情况。

13.15 参见习题13.14。

a. 使用对数连结拟合负二项模型。对结果加以解释。将计数与壳宽进行绘图,并指出哪一个连结函数更为合适。

b. 利用壳宽作为预测变量,使用对数连结拟合泊松广义线性混合模型。对结果加以解释。

c. 比较各种模型的结果,包括第4.3.2节中泊松广义线性模型的结果。指出你倾向于选择哪一个模型,并说明理由。

13.16 参考习题13.14和13.15。利用壳宽和颜色的类别作为预测变量,拟合:(a)负二项广义线性模型;(b)泊松广义线性混合模型。检查是否存在交互效应,并解释

最终的模型。

13.17 参见表 13.6。对于那些"其他"种族的被访者,报告认识的人中有(0,1,2,3,4,5,6)个凶杀受害者的样本计数分别为(55,5,1,0,1,0,0)。对这些数据以及有关白人和黑人的数据同时拟合一个合适的模型。比较每两对之间均值的差异,对模型结果加以解释。

13.18 利用类似然法分析表 13.6 中有关凶杀受害者的计数。

13.19 使用负二项广义线性模型,对习题 4.6 中有关电脑芯片生产中的缺陷数据进行分析。将结果与泊松广义线性模型的相应结果进行比较。指出为什么两个模型的结果很相似。

13.20 根据本书网站(*www.stat.ufl.edu/~aa/cda/cda.html*)上的数据,利用本章所介绍的方法分析在 2000 年总统大选中改革党候选人 Pat Buchanan 与 1996 年总统大选中该党候选人 Ross Perot 相比所获得的全国选票的情况。注意棕榈滩县是一个非常大的奇异值(outlier)(很显然,这主要是由于选票的迷惑性导致许多本来投给 Al Gore 的选票被认定为投给了 Buchanan)。在分别包括和剔除该观测值的情况下进行模型分析,并对结果加以比较。

13.21 构建一个潜类模型,分析 Espeland 和 Handelman(1989)的数据。

13.22 参考 Liang 和 Hanfelt(1994)的畸形学研究数据。利用至少两种不同的处理过度离散的二分数据的方法来分析这些数据。比较有关结果,并加以解释。

13.23 参考习题 13.14。从壳宽、体重、颜色,以及刺毛状况中选取合适的变量作为预测变量,找出一个预测同伴数量的合理模型,并加以解释。

理论与方法

13.24 推导具有 q 个潜类的潜类模型的残差自由度。当 $I=2$ 时,证明在 $q\geq 2$ 的情况下,需要 $T\geq 4$ 才使得模型为非饱和模型。给出当 $T=4,5$ 时 q 的最大取值。对于一个 I^2 表格,证明需要 $q<I^2/(2I-1)$。

13.25 利用模型的参数表示式 13.1 潜类模型的对数似然函数,并推导它的似然方程(Goodman, 1974; Haberman, 1979)。

13.26 令 Π 表示一个关于 X 和 Y 的联合分布单元格概率的矩阵。假定存在有关概率的 $I\times 1$ 列向量 $\boldsymbol{\pi}_{1k}$ 和 $J\times 1$ 列向量 $\boldsymbol{\pi}_{2k}(k=1,\cdots,q)$,以及一组概率 $\{\rho_k\}$ 使得

$$\Pi = \sum_{k=1}^{q}\rho_k\boldsymbol{\pi}_{1k}\boldsymbol{\pi}'_{2k}。$$

说明为什么这表明存在一个潜变量 Z,使得在给定 Z 后 X 与 Y 条件独立。

13.27 在第 13.2.2 节中,在普通 logistic 回归模型成立的零假设下,说明为什么将该模型的偏离度与由两个 logistic 回归所形成的混合模型的偏离度之差作为卡方统计量是不适当的。

13.28 参考习题 12.7。令 $\mu_k(a,b,c)$ 表示在第 k 种干预次序下干预方式(A,B,C)所对应的结果(a,b,c)的期望频数,其中结果 1 = 好转,0 = 没有好转。利用非参数随机效应方法,表明通过拟合下列准对称性模型可以估计式 12.19 模型中的干预效应:

$$\log\mu_k(a,b,c) = a\beta_A + b\beta_B + c\beta_C + \lambda_k(a,b,c),$$

其中,$\lambda_k(a,b,c)=\lambda_k(a,c,b)=\lambda_k(b,a,c)=\lambda_k(b,c,a)=\lambda_k(c,a,b)=\lambda_k(c,b,a)$。拟合该模型,并展示 $\hat{\beta}_B-\hat{\beta}_A=1.64$(SE = 0.34),$\hat{\beta}_C-\hat{\beta}_A=2.23$(SE = 0.39),$\hat{\beta}_C-\hat{\beta}_B=0.59$(SE = 0.39)。对这些结果加以解释,并将其与习题 12.7 中式12.19模型的结果进行对比。

13.29 证明当 $\theta=0$ 时,β-二项分布(式 13.9)简化为二项分布。

13.30 将β密度函数的分子表示为关于μ和θ的函数。据此证明:(a)当$\theta < \min(\mu, 1-\mu)$时,该函数是单峰的;(b)当$\mu = \theta = \frac{1}{2}$时,它等于均匀密度函数。

13.31 假定$\pi_i = P(Y_{it} = 1) = 1 - P(Y_{it} = 0)$,其中$t = 1, \cdots, n_i$,并且对于$t \neq s$,$\text{corr}(Y_{it}, Y_{is}) = \rho$。证明$\text{var}(Y_{it}) = \pi_i(1-\pi_i)$,$\text{cov}(Y_{it}, Y_{is}) = \rho\pi_i(1-\pi_i)$,以及

$$\text{var}\left(\sum_t Y_{it}\right) = n_i\pi_i(1-\pi_i)\left[1 + \rho(n_i - 1)\right]。$$

13.32 当$n = 1$时,证明β-二项分布与二项分布(即伯努利分布)没有区别。说明为什么当$n = 1$时不可能发生过度离散问题。

13.33 当y_i等于n_i个均值分别为μ_i的二分结果变量之和时,参考具有$v(\mu_i) = \phi n_i \mu_i (1-\mu_i)$的类似然法,说明为什么这个方差函数存在结构性问题:当$n_i = 1$时,只有$\phi = 1$才合理。

13.34 Liang 和 Hanfelt(1994)介绍了一项对控制组和干预组进行比较的畸形学研究,其中对应于假定与不假定每组中的离散参数相同两种情形,β-二项分布模型对干预效应的最大似然估计相差两倍。相反,无论假定每组中的ϕ相同与否,在具有式13.11方差函数的类似然法中对干预效应的估计都相等。解释原因,并讨论这是该方法的优点还是缺点。

13.35 考虑 logistic-正态模型,$\text{Logit}(\pi_i) = \alpha + \mathbf{x}_i'\boldsymbol{\beta} + u_i$。在$\sigma$取值很小的情况下,证明它近似于一个混合模型,其中混合分布具有$\text{var}(\pi_i) = [\mu_i(1-\mu_i)]^2\sigma^2$(提示:参考习题6.33)。

13.36 Altham(1978)介绍了离散分布

$$f(y;\pi,\psi) = c(\pi,\psi)\binom{n}{y}\pi^y(1-\pi)^{n-y}\exp\left[\psi y(n-y)\right]$$

$$y = 0, 1, \cdots, n,$$

其中,$c(\pi,\psi)$是一个正态化的常数。证明该分布属于指数族分布。证明当$\psi = 0$时,它等于二项分布(Altham 指出,当$\psi < 0$时,会出现过度离散的问题。Corcoran 等(2001)以及 Lindsey 和 Altham(1998)以此为基础,提出了除β-二项分布模型之外的另一种模型)。

13.37 在k值固定的负二项分布(式13.13)中,$y_1, \cdots, y_N$相互独立,证明$\hat{\mu} = \bar{y}$。

13.38 利用$E(Y) = E[E(Y|X)]$以及$\text{var}(Y) = E[\text{var}(Y|X)] + \text{var}[E(Y|X)]$,推导下列分布的均值和方差:(a)$\beta$-二项分布;(b)负二项分布。

13.39 假定在给定u后,Y服从$E(Y|u) = u\mu$的泊松分布,其中μ的取值取决于预测变量。假定u是一个正的随机变量,并有$E(u) = 1$且$var(u) = \tau$,证明$E(Y) = \mu$且$\text{var}(Y) = \mu + \tau\mu^2$。说明为什么使用对数连结的负二项广义线性模型和泊松广义线性混合模型都是它的特例。

13.40 负二项分布的另一种参数化形式来自于γ分布的密度公式,

$$f(\lambda; k, \mu) = \frac{(k)^{k\mu}}{\Gamma(k\mu)}\exp(-k\lambda)\lambda^{k\mu-1}, \lambda \geq 0,$$

其中,$E(\lambda) = \mu$,$\text{var}(\lambda) = \mu/k$。证明关于泊松分布的$\gamma$混合可推导出一个具有

$$E(Y) = \mu, \text{var}(Y) = \mu(1+k)/k$$

的负二项分布。当k取什么极限值时,这个分布简化为泊松分布?(有关的最大似然模型拟合,参见:Nelder and Lee(1996)。Cameron and Trivedi(1998, p. 75)指出,与二次项方差函数不同,当关于均值的模型成立但真实分布不服从负二项

分布时,参数的估计值不具有一致性)。

13.41　负二项分布是一个单峰分布,它的众数等于 $\mu(k-1)/k$ 的整数部分(Johnson et al., 1992, pp. 208-209)。证明当 $\mu \leq 1$ 时,它的众数等于0;当 $\mu > 1$ 但 $k < \mu/(\mu-1)$ 时,它的众数仍然为0(由此推出,它的斜率大于泊松分布的斜率,因为后者的众数等于其均值的整数部分)。

13.42　考虑对数线性随机效应模型

$$\log[E(Y_{it} \mid \mathbf{u}_i)] = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i,$$

其中,$\{\mathbf{u}_i\}$ 服从独立的 $N(0, \boldsymbol{\Sigma})$ 分布。证明它在边际上对应的对数线性模型为

$$\log[E(Y_{it})] - \frac{1}{2}\mathbf{z}'_{it}\boldsymbol{\Sigma}\mathbf{z}_{it} = \mathbf{x}'_{it}\boldsymbol{\beta},$$

其中,固定效应相同,但却包括了一个抵消项。在只存在随机截距的情况下,说明 σ 对于抵消项大小的影响。解释当 $\sigma=0$ 时,会出现什么情况。

13.43　在第13.5.1节和习题13.42中,我们看到对于泊松广义线性混合模型来说,边际效应与群组别效应相同。这并不意味着泊松广义线性混合模型和泊松广义线性模型的效应的最大似然估计也相同。说明为什么(提示:在广义线性混合模型中,边际分布服从泊松分布吗?)。

13.44　对于泊松广义线性混合模型(式13.14),利用正态矩量生成函数(mgf)证明,当 $t \neq s$ 时,

$$\text{cov}(Y_{it}, Y_{is}) = \exp[(\mathbf{x}'_{it} + \mathbf{x}'_{is})\boldsymbol{\beta}][\exp(\sigma^2)(\exp(\sigma^2) - 1)]$$

并求解 $\text{corr}(Y_{it}, Y_{is})$。

13.45　考虑使用恒等连结的泊松广义线性混合模型。给出边际均值和方差与相应的条件均值和方差的关系。指出这个模型所存在的结构性问题。

14 参数模型的渐近理论

与其他章节相比，本章侧重探讨理论问题。在这里，我们介绍关于分类数据参数模型的渐近理论，重点讨论有关列联表的多项分布模型。在第 14.1 节，我们回顾 δ 方法并对其进行扩展。该方法被用于推导许多统计量的大样本正态分布。在第 14.2 节，我们应用 δ 方法推导列联表模型的最大似然参数估计，并在第 14.4 节介绍有关 logit 模型和对数线性模型的情况。在第 14.3 节，我们推导关于单元格残差以及 X^2 和 G^2 拟合优度统计量的渐近分布。

本章介绍的内容可谓历史悠久。Pearson(1900)推导出了检验某一特定多项分布的 X^2 统计量渐近服从卡方分布。Fisher(1922,1924)给出了当多项分布概率为未知参数的函数时对自由度所进行的调整。Cramér(1946, pp. 424-434)在假设参数的最大似然估计值具有一致性的条件下，对此结果进行了正式的证明。Rao(1957)则证明了在一般条件下最大似然估计值满足一致性。尽管 Rao 的研究关注的重点是对一致性的证明，他同时也给出了最大似然估计值的渐近分布。Birch(1964a)证明了这些结果在更弱的条件下也成立。Anderson(1980)、Bishop 等(1975)、Cox(1984)、Haberman(1974a)以及 Watson(1959)都在此方面作出了重要贡献。

与 Cramér 和 Rao 的证明一致，我们的推导也将最大似然估计值视为在参数空间中对数似然函数的导数等于零的点。Birch 将最大似然估计值当做似然函数在该处的取值无限接近于其上确界(supremum)的点。尽管 Birch 的方法更强大，但是证明起来也更加复杂。在此，我们避免使用“定理-证明”式的正式表述方式。相反，我们将展示，这些重要结果实际上来源于一些简单的数学思想，比如泰勒级数展开。

14.1 δ 方法

假定估计参数的统计量服从大样本正态分布。在本节我们将表明，在这种情况下，这些统计量的许多函数也渐近服从正态分布。

14.1.1 O，o 收敛率

在描述一系列极限的时候，O 和 o 的区分很重要。对于实数 $\{z_n\}$，$o(z_n)$ 表示在 $n\to\infty$ 时它比 z_n 的阶数更小，即随着 $n\to\infty$，存在 $o(z_n)/z_n\to 0$。举例来说，在 $n\to\infty$ 时，$\sqrt{n}$ 就可以表示为 $o(n)$，因为当 $n\to\infty$ 时，$\sqrt{n}/n\to 0$。如果序列满足 $o(1)/1=o(1)\to 0$，我们称之为 $o(1)$；例如，当 $n\to\infty$ 时，$n^{-1/2}$ 就可以表示为 $o(1)$。

$O(z_n)$表示与z_n同阶的项，即随着$n \to \infty$，$|O(z_n)/z_n|$是有边界的。例如，当$n \to \infty$时，$(3/n)+(8/n^2)$就可以表示为$O(n^{-1})$，因为随着n不断增大，它与n^{-1}的比近似等于3。

类似的标识也可以用来表示一组随机变量。在标识中加上下标p，用来表明这组变量的关系是从概率意义上讲的，而不是绝对意义上的。这样，符号$o_p(z_n)$代表在n很大的情况下，阶数小于z_n的随机变量，即$o_p(z_n)/z_n$*依概率收敛于*（*converges in probability*）0；也即，对于任意给定的$\varepsilon>0$，随着$n \to \infty$，$P(|o_p(z_n)/z_n| \leqslant \varepsilon) \to 1$。相反，符号$O_p(z_n)$代表这样一个随机变量，对于任意$\varepsilon>0$，都存在常数$K$和整数$n_0$，使得对所有$n>n_0$都满足$P[|O_p(z_n)/z_n|<K]>1-\varepsilon$。

具体来说，令$\bar{Y}_n$表示服从某一具有$E(Y_i)=\mu$的分布的n个独立观测值$Y_1,\cdots,Y_n$的样本均值。那么$(\bar{Y}_n-\mu)=o_p(1)$，因为根据大数定理，随着$n \to \infty$，$(\bar{Y}_n-\mu)/1$依概率收敛于零。按照切比雪夫不等式（Tchebychev's inequality），一个随机变量和它的期望值之差与该随机变量的标准差是同阶的。由于$\bar{Y}_n-\mu$的标准差为$\sigma/\sqrt{n}$，所以$(\bar{Y}_n-\mu)=O_p(n^{-1/2})$。

可以表示为$O_p(n^{-1/2})$的随机变量同时也满足$o_p(1)$。这样的一个例子就是$(\bar{Y}_n-\mu)$。乘法运算对阶数的影响与普通运算中的情形相同（习题14.1）。例如，$\sqrt{n}(\bar{Y}_n-\mu)=n^{1/2}O_p(n^{-1/2})=O_p(n^{1/2}n^{-1/2})=O_p(1)$。如果随着$n \to \infty$，两个随机变量之差满足$o_p(1)$，斯拉茨基定理（Slutzky's theorem）表明，这些随机变量的极限具有相同的分布。

14.1.2　随机变量函数的δ方法

令T_n表示一个统计量，其下标表明该统计量的取值依赖于样本规模n。在大样本的情况下，假定T_n在θ附近近似服从正态分布，其标准误约等于$\sigma/\sqrt{n}$。更确切地说，随着$n \to \infty$，假定$\sqrt{n}(T_n-\theta)$的累积分布函数收敛于$N(0,\sigma^2)$的累积分布函数。这个极限是一个*依分布收敛*（*convergence in distribution*）的例子，可表示为

$$\sqrt{n}(T_n-\theta) \xrightarrow{d} N(0,\sigma^2)。\qquad (14.1)$$

接下来，我们对函数g推导关于$g(T_n)$的极限分布。假定g在θ处至少是二阶可导的，应用$g(t)$在θ附近的泰勒级数展开，对t和θ之间的某个θ^*存在：

$$\begin{aligned} g(t) &= g(\theta)+(t-\theta)g'(\theta)+(t-\theta)^2 g''(\theta^*)/2 \\ &= g(\theta)+(t-\theta)g'(\theta)+O(|t-\theta|^2)。\end{aligned}$$

对应于上式中的t，代入随机变量T_n，可以得出

$$\begin{aligned} \sqrt{n}[g(T_n)-g(\theta)] &= \sqrt{n}(T_n-\theta)g'(\theta)+\sqrt{n}O(|T_n-\theta|^2) \\ &= \sqrt{n}(T_n-\theta)g'(\theta)+O_p(n^{-1/2}) \end{aligned} \qquad (14.2)$$

其中

$$\sqrt{n}O(|T_n-\theta|^2)=\sqrt{n}O[O_p(n^{-1})]=O_p(n^{-1/2})。$$

由于$O_p(n^{-1/2})$项在渐近过程中可以忽略，$\sqrt{n}[g(T_n)-g(\theta)]$具有与$\sqrt{n}(T_n-\theta)g'(\theta)$相同的极限分布，也即，$g(T_n)-g(\theta)$相当于对$(T_n-\theta)$乘以一个常数$g'(\theta)$。由于$(T_n-\theta)$近似服从方差为$\sigma^2/n$的正态分布，因此，$g(T_n)-g(\theta)$近似服从方差为$\sigma^2[g'(\theta)]^2/n$的正态分布。更确切地说，

$$\sqrt{n}[g(T_n)-g(\theta)] \xrightarrow{d} N(0,\sigma^2[g'(\theta)]^2)。\qquad (14.3)$$

图 3.1 显示了这一结果,并且在第 3.1.6 节中我们将这一结果应用于样本 logit。

结果(式 14.3)被称为获得渐近分布的 δ 方法(*delta method*)。由于 $\sigma^2=\sigma^2(\theta)$ 并且 $g'(\theta)$ 一般也取决于 θ,渐近方差是未知的。令 $\sigma^2(T_n)$ 和 $g'(T_n)$ 表示关于在 θ 的样本估计值 T_n 时的相应取值。当 $g'(\cdot)$ 和 $\sigma=\sigma(\cdot)$ 在 θ 处连续时,$\sigma(T_n)g'(T_n)$ 是关于 $\sigma(\theta)g'(\theta)$ 的一致性估计。因此,置信区间以及假设检验利用了 $\sqrt{n}[g(T_n)-g(\theta)]/\sigma(T_n)|g'(T_n)|$ 渐近服从标准正态分布这一结果。例如,

$$g(T_n)\pm z_{\alpha/2}\sigma(T_n)|g'(T_n)|/\sqrt{n}$$

是关于 $g(\theta)$ 的 $100(1-\alpha)\%$ 的大样本置信区间。

当 $g'(\theta)=0$ 时,式 14.3 没有意义,因为它的极限方差等于 0。在这种情况下,$\sqrt{n}[g(T_n)-g(\theta)]=o_p(1)$,而泰勒级数展开中的高阶项给出了它的渐近分布(参见注解14.1)。

14.1.3 随机向量函数的 δ 方法

δ 方法可以扩展到随机向量函数的情况。假定 $\mathbf{T}_n=(T_{n1},\cdots,T_{nN})'$ 渐近服从多元正态分布,其中均值 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_N)'$,协方差矩阵为 $\boldsymbol{\Sigma}/n$。假设 $g(t_1,\cdots,t_N)$ 在 $\boldsymbol{\theta}$ 处具有非零的导数 $\boldsymbol{\phi}=(\phi_1,\cdots,\phi_N)'$,其中

$$\phi_i=\frac{\partial g}{\partial t_i}\Big|_{t=\theta},$$

那么,

$$\sqrt{n}[g(\mathbf{T}_n)-g(\boldsymbol{\theta})]\xrightarrow{d}N(\mathbf{0},\boldsymbol{\phi}'\boldsymbol{\Sigma}\boldsymbol{\phi})。\tag{14.4}$$

在大样本的情况下,$g(\mathbf{T}_n)$ 的分布与均值为 $g(\boldsymbol{\theta})$、方差为 $\boldsymbol{\phi}'\boldsymbol{\Sigma}\boldsymbol{\phi}/n$ 的正态分布相似。

对式 14.4 的证明可以由下列展开式推得:

$$g(\mathbf{T}_n)-g(\boldsymbol{\theta})=(\mathbf{T}_n-\boldsymbol{\theta})'\boldsymbol{\phi}+o(\|\mathbf{T}_n-\boldsymbol{\theta}\|),$$

其中,$\|\mathbf{z}\|=(\sum z_i^2)^{1/2}$ 代表向量 $\mathbf{z}$ 的长度。在 n 很大的情况下,$g(\mathbf{T}_n)-g(\boldsymbol{\theta})$ 相当于随机向量 $(\mathbf{T}_n-\boldsymbol{\theta})$ 的一个线性函数,而 $(\mathbf{T}_n-\boldsymbol{\theta})$ 近似服从正态分布。因而,$g(\mathbf{T}_n)-g(\boldsymbol{\theta})$ 也近似服从正态分布。

14.1.4 多项分布计数函数的渐近正态性

关于随机向量的 δ 方法表明,列联表单元格计数的许多函数也具有渐近正态性。假定单元格计数 $(n_1,\cdots,n_N)$ 服从单元格概率为 $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_N)'$ 的多项分布,令 $n=n_1+\cdots+n_N$,并令 $\mathbf{p}=(p_1,\cdots,p_N)'$ 代表相应的样本比例,其中 $p_i=n_i/n$。

将列联表交叉划分的 n 个观测值中的第 i 个表示为 $\mathbf{Y}_i=(Y_{i1},\cdots,Y_{iN})$,如果它落在第 j 个单元格,$Y_{ij}=1$,否则 $Y_{ij}=0$,其中 $i=1,\cdots,n$。例如,$\mathbf{Y}_6=(0,0,1,0,0,\cdots,0)$ 表明,第 6 个观测值落在表中的第三个单元格。由于每个观测值只能落在一个单元格中,因而 $\sum_j Y_{ij}=1$,并且当 $j\neq k$ 时存在 $Y_{ij}Y_{ik}=0$。同时,$p_j=\sum_i Y_{ij}/n$,并且 $E(Y_{ij})=P(Y_{ij}=1)=\pi_j=E(Y_{ij}^2)$,如果 $j\neq k$,$E(Y_{ij}Y_{ik})=0$。

由此推出

$$E(\mathbf{Y}_i)=\boldsymbol{\pi}\text{ 以及 }\operatorname{cov}(\mathbf{Y}_i)=\boldsymbol{\Sigma},\quad i=1,\cdots,n,$$

其中 $\boldsymbol{\Sigma}=(\sigma_{jk})$,有:

$$\sigma_{jj}=\operatorname{var}(Y_{ij})=E(Y_{ij}^2)-[E(Y_{ij})]^2=\pi_j(1-\pi_j),$$

对于 $j \neq k$，$\sigma_{jk} = \text{cov}(Y_{ij}, Y_{ik}) = E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik}) = -\pi_j\pi_k$。
矩阵 $\boldsymbol{\Sigma}$ 具有如下形式：

$$\boldsymbol{\Sigma} = \textbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi\pi}',$$

这里，$\textbf{diag}(\boldsymbol{\pi})$ 是主对角线元素等于 π 的对角矩阵。

由于 $\mathbf{p}$ 是 n 个独立观测值的样本均值，即

$$\mathbf{p} = \frac{\sum_{i=1}^{n} \mathbf{Y}_i}{n},$$

$$\text{cov}(\mathbf{p}) = [\textbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi\pi}']/n。 \tag{14.5}$$

因为存在线性相依关系 $\sum p_i = 1$，这个协方差矩阵是一个奇异矩阵。根据多元的中心极限定理(Rao,1973,p.128)，

$$\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{d} N[\mathbf{0}, \textbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi\pi}']。 \tag{14.6}$$

按照 δ 方法，在 $\boldsymbol{\pi}$ 处存在非零导数的 $\mathbf{p}$ 的函数也渐近服从正态分布。令 $g(t_1, \cdots t_N)$ 表示一个可导函数，并令

$$\phi_i = \partial g / \partial \pi_i, \quad i = 1, \cdots, N,$$

表示 $\partial g/\partial t_i$ 在 $\mathbf{t} = \boldsymbol{\pi}$ 时的取值。利用 δ 方法(式 14.4)有，

$$\sqrt{n}[g(\mathbf{p}) - g(\boldsymbol{\pi})] \xrightarrow{d} N(0, \boldsymbol{\phi}'[\textbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi\pi}']\boldsymbol{\phi})。 \tag{14.7}$$

其渐近方差等于

$$\boldsymbol{\phi}'\textbf{diag}(\boldsymbol{\pi})\boldsymbol{\phi} - (\boldsymbol{\phi}'\boldsymbol{\pi})^2 = \sum \pi_i\phi_i^2 - (\sum \pi_i\phi_i)^2。$$

在第 3.1.7 节中，我们在推导样本对数发生比之比的大样本方差时曾用到该公式。

14.1.5　随机向量的向量函数的 δ 方法

δ 方法可以进一步扩展到有关渐近正态随机向量的向量函数的情况。令 $g(\mathbf{t}) = (g_1(\mathbf{t}), \cdots, g_q(\mathbf{t}))'$ 并令 $(\partial \mathbf{g}/\partial \boldsymbol{\theta})$ 表示 $q \times N$ 雅可比(Jacobian)矩阵，其第 i 行和第 j 列的元素为 $\partial g_i(\mathbf{t})/\partial t_j$ 在 $\mathbf{t} = \boldsymbol{\theta}$ 处的取值，那么，

$$\sqrt{n}[\mathbf{g}(\mathbf{T}_n) - g(\boldsymbol{\theta})] \xrightarrow{d} N[\mathbf{0}, (\partial \mathbf{g}/\partial \boldsymbol{\theta})\boldsymbol{\Sigma}(\partial \mathbf{g}/\partial \boldsymbol{\theta})']。 \tag{14.8}$$

极限正态分布的秩等于相应渐近协方差矩阵的秩。

式 14.8 在求解大样本联合分布时非常有用。例如，根据式 14.6、式 14.7 以及式14.8，多项分布比例的几个函数的渐近分布具有以下形式的协方差矩阵：

$$\text{asymp. cov}\{\sqrt{n}[\mathbf{g}(\mathbf{p}) - \mathbf{g}(\boldsymbol{\pi})]\} = \boldsymbol{\Phi}[\textbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi\pi}']\boldsymbol{\Phi}',$$

其中，$\boldsymbol{\Phi}$ 是指雅可比矩阵 $(\partial \mathbf{g}/\partial \boldsymbol{\pi})$。

14.1.6　对数发生比之比的联合渐近正态性

我们通过求解列联表中的一组对数发生比之比的联合渐近分布，对公式 14.8 进行具体说明。这里，我们使用对数尺度，因为它对正态性的收敛速度更快。

令 $\mathbf{g}(\boldsymbol{\pi}) = \log(\boldsymbol{\pi})$ 表示由单元格概率的自然对数组成的向量，其中

$$\partial \mathbf{g}/\partial \boldsymbol{\pi} = \textbf{diag}(\boldsymbol{\pi})^{-1}。$$

$\sqrt{n}[\log(\mathbf{p}) - \log(\boldsymbol{\pi})]$ 的渐近分布的协方差等于

$$\text{diag}(\boldsymbol{\pi})^{-1}[\textbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi\pi}']\textbf{diag}(\boldsymbol{\pi})^{-1} = \textbf{diag}(\boldsymbol{\pi})^{-1} - \mathbf{11}'$$

其中，$\mathbf{1}$ 代表一个元素为 1 的 $N \times 1$ 向量。

对于一个 $q \times N$ 的常数矩阵 $\mathbf{C}$,可推出:

$$\sqrt{n}\mathbf{C}[\log(\mathbf{p}) - \log(\boldsymbol{\pi})] \xrightarrow{d} N[\mathbf{0}, \mathbf{C}\mathbf{diag}(\boldsymbol{\pi})^{-1}\mathbf{C}' - \mathbf{C}\mathbf{1}\mathbf{1}'\mathbf{C}']。 \quad (14.9)$$

这里,假定 $\mathbf{C}\log(\mathbf{p})$ 是一组样本对数发生比之比,那么,$\mathbf{C}$ 的每一行中,除了与形成给定对数发生比之比的 $\log(\mathbf{p})$ 所对应的位置元素为两个 $+1$ 和两个 -1 外,其他元素均等于零。这时,式 14.9 中的协方差矩阵的第二项也等于零。如果某个特定的发生比之比由单元格 h,i,j 和 k 组成,相应渐近分布的方差为

$$\text{asymp. var}[\sqrt{n}(\text{样本对数发生比之比})] = \pi_h^{-1} + \pi_i^{-1} + \pi_j^{-1} + \pi_k^{-1}。$$

当两个对数发生比之比所使用的单元格都互不相同时,在极限分布中它们之间的渐近协方差等于零。

14.2 模型参数和单元格概率估计值的渐近分布

现在,我们推导在大样本情况下关于列联表的模型推断的基本结果。δ 方法是我们所使用的主要工具,相应推导适用于单一的多项分布。当参数空间随着样本规模的增加保持固定不变时,这些结果可以直接扩展到多个多项分布乘积的情况。

在一个包括 N 个单元格的列联表中,观测值为单元格计数 $\mathbf{n} = (n_1, \cdots, n_N)'$。渐近过程将 N 视为给定的,并令 $n = \sum n_i \to \infty$。我们假定 $\mathbf{n} = n\mathbf{p}$ 服从概率为 $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_N)'$ 的多项分布。模型为

$$\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta}),$$

其中 $\boldsymbol{\pi}(\boldsymbol{\theta})$ 将 $\boldsymbol{\pi}$ 表示为关于一组数量较少的参数 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_q)'$ 的函数。

随着 $\boldsymbol{\theta}$ 的取值在其参数空间内变动,$\boldsymbol{\pi}(\boldsymbol{\theta})$ 的取值落在 N 个概率 $\boldsymbol{\pi}$ 的一个子空间里。当在 $\boldsymbol{\theta}$ 中加入新的项时,模型变得更加复杂,进而满足模型的 $\boldsymbol{\pi}$ 的空间也变得更大。利用 $\boldsymbol{\theta}$ 和 $\boldsymbol{\pi}$ 表示具有一般性的参数和概率值,$\boldsymbol{\theta}_0 = (\theta_{10}, \cdots, \theta_{q0})'$ 和 $\boldsymbol{\pi}_0 = (\pi_{10}, \cdots, \pi_{N0})' = \boldsymbol{\pi}(\boldsymbol{\theta}_0)$ 表示某个特定的真值。当模型不成立时,满足 $\boldsymbol{\pi}(\boldsymbol{\theta}_0) = \boldsymbol{\pi}_0$ 的 $\boldsymbol{\theta}_0$ 不存在;也即,对于参数空间内所有可能的 $\boldsymbol{\theta}$,$\boldsymbol{\pi}_0$ 落在了 $\boldsymbol{\pi}(\boldsymbol{\theta})$ 的取值范围 $\boldsymbol{\pi}$ 之外。这种情况我们将在第 14.3.5节讨论。

首先,我们推导关于 $\boldsymbol{\theta}$ 的最大似然估计值 $\hat{\boldsymbol{\theta}}$ 的渐近分布。利用该结果,可以继续推导关于 $\boldsymbol{\pi}$ 的模型最大似然估计值 $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ 的渐近分布。这里借鉴了 Rao(1973, Sec. 5e) 以及 Bishop 等(1975, Secs. 14.7 and 14.8)的方法,相应方法假定满足下列规则性条件:

1. $\boldsymbol{\theta}_0$ 没有落在参数空间的边界上。
2. 所有的 $\pi_{i0} > 0$。
3. $\boldsymbol{\pi}(\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}_0$ 附近存在连续的一阶偏导数。
4. 雅可比矩阵($\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}$)在 $\boldsymbol{\theta}_0$ 处满秩,其秩为 q。

这些条件确保 $\boldsymbol{\pi}(\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}_0$ 处局部平滑,并且为一一对应的函数,在 $\boldsymbol{\theta}_0$ 和 $\boldsymbol{\pi}_0$ 附近可以进行泰勒级数展开。当雅可比矩阵不满秩时,一般可以通过重新构建具有更少参数的模型来实现这一条件。

14.2.1 模型参数估计值的渐近分布

对 $\hat{\boldsymbol{\theta}}$ 的分布进行推导的关键在于,将 $\hat{\boldsymbol{\theta}}$ 表示为一个关于 $\mathbf{p}$ 的线性函数。这时,就可以根据 $\mathbf{p}$ 的渐近正态性应用 δ 方法。这个线性化过程包括两步:首先将 $\mathbf{p}$ 与 $\hat{\boldsymbol{\pi}}$ 建立联

系，然后将 $\hat{\boldsymbol{\pi}}$ 与 $\hat{\boldsymbol{\theta}}$ 相联系。

多项分布对数似然函数的核函数为

$$L(\boldsymbol{\theta}) = \log \prod_{i=1}^{N} \pi_i(\boldsymbol{\theta})^{n_i} = n\sum_{i=1}^{N} p_i \log \pi_i(\boldsymbol{\theta})。$$

相应的似然方程为

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_j} = n\sum_i \frac{p_i}{\pi_i(\boldsymbol{\theta})}\frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1,\cdots,q。 \tag{14.10}$$

这些方程取决于模型中所选择的 $\boldsymbol{\pi}(\boldsymbol{\theta})$ 的函数形式。注意

$$\sum_i \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\left[\sum_i \pi_i(\boldsymbol{\theta})\right] = \frac{\partial}{\partial \theta_j}(1) = 0。 \tag{14.11}$$

令 $\partial\pi_i/\partial\hat{\theta}_j$ 表示 $\partial\pi_i(\boldsymbol{\theta})/\partial\theta_j$ 在 $\hat{\boldsymbol{\theta}}$ 处的取值。从第 j 个式 14.10 似然方程的两边同时消去相同的项，有

$$\sum_i \frac{n(p_i - \pi_{i0})}{\hat{\pi}_i}\frac{\partial \pi_i}{\partial \hat{\theta}_j} = \sum_i \frac{n(\hat{\pi}_i - \pi_{i0})}{\hat{\pi}_i}\frac{\partial \pi_i}{\partial \hat{\theta}_j}, \tag{14.12}$$

这里根据式 14.11，方程右边的第一项求和等于零。

接下来，我们通过

$$\hat{\pi}_i - \pi_{i0} = \sum_k (\hat{\theta}_k - \theta_{k0})\frac{\partial \pi_i}{\partial \bar{\theta}_k}$$

将 $\hat{\boldsymbol{\pi}}$ 表示为 $\hat{\boldsymbol{\theta}}$ 的函数，其中 $\partial\pi_i/\partial\bar{\theta}_k$ 代表 $\partial\pi_i/\partial\boldsymbol{\theta}_k$ 在 $\hat{\boldsymbol{\theta}}$ 和 $\boldsymbol{\theta}_0$ 之间的某一点 $\bar{\boldsymbol{\theta}}$ 处的取值。将其代入等式 14.12 的右边，并将两边同除以 $\sqrt{n}$ 得出，对于每个 j，

$$\sum_i \frac{\sqrt{n}(p_i - \pi_{i0})}{\hat{\pi}_i}\frac{\partial \pi_i}{\partial \hat{\theta}_j} = \sum_k \sqrt{n}(\hat{\theta}_k - \theta_{k0})\left(\sum_i \frac{1}{\hat{\pi}_i}\frac{\partial \pi_i}{\partial \hat{\theta}_j}\frac{\partial \pi_i}{\partial \bar{\theta}_k}\right)。 \tag{14.13}$$

利用一些符号可以更简便地表述 $\hat{\boldsymbol{\theta}}$ 对 $\mathbf{p}$ 的依赖关系。令 $\mathbf{A}$ 表示一个 $N \times q$ 矩阵，其元素为

$$a_{ij} = \pi_{i0}^{-1/2}\frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_{j0}}。$$

A 的矩阵表达形式为

$$\mathbf{A} = \mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0), \tag{14.14}$$

其中 $(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)$ 代表雅可比矩阵 $(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}_0$ 处的取值。随着 $\hat{\boldsymbol{\theta}}$ 收敛于 $\boldsymbol{\theta}_0$，式 14.13 右边括号中的项收敛于 $\mathbf{A'A}$ 在第 j 行、第 k 列的元素。随着 $\hat{\boldsymbol{\theta}}\to\boldsymbol{\theta}_0$，方程组 14.13 可表示为

$$\mathbf{A'diag}(\boldsymbol{\pi}_0)^{-1/2}\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}_0) = (\mathbf{A'A})\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1)。$$

由于雅可比矩阵在 $\boldsymbol{\theta}_0$ 处满秩，$\mathbf{A'A}$ 是非奇异矩阵，因此，

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = (\mathbf{A'A})^{-1}\mathbf{A'diag}(\boldsymbol{\pi}_0)^{-1/2}\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}_0) + o_p(1)。 \tag{14.15}$$

这里，$\mathbf{p}$ 的渐近分布决定了 $\hat{\boldsymbol{\theta}}$ 的渐近分布。由式 14.6 可得，$\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}_0)$ 渐近服从正态分布，其协方差矩阵为 $[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}'_0]$。根据 δ 方法，$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ 也渐近服从正态分布，它的渐近协方差矩阵为

$$(\mathbf{A'A})^{-1}\mathbf{A'diag}(\boldsymbol{\pi}_0)^{-1/2} \times [\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}'_0] \times \mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}\mathbf{A}(\mathbf{A'A})^{-1}。$$

利用式 14.11 和式 14.14，可以消去上式中的相减项，因为

$$\boldsymbol{\pi}'_0\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}\mathbf{A} = \boldsymbol{\pi}'_0\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)$$
$$= \mathbf{1}'(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0) = \left(\sum_i \partial\pi_i/\partial\boldsymbol{\theta}_0\right)' = \mathbf{0}'。$$

因此，$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$的渐近协方差矩阵简化为$(\mathbf{A}'\mathbf{A})^{-1}$。

总之，这个推导过程给出了一般性的结论

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}N[\mathbf{0},(\mathbf{A}'\mathbf{A})^{-1}]。\tag{14.16}$$

$\hat{\boldsymbol{\theta}}$的渐近协方差矩阵取决于$(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)$，进而取决于模型中$\boldsymbol{\pi}$作为$\boldsymbol{\theta}$的函数的形式。令$\hat{\mathbf{A}}$表示$\mathbf{A}$在最大似然估计值$\hat{\boldsymbol{\theta}}$处的取值，所估计的协方差矩阵等于

$$\widehat{\text{cov}}(\hat{\boldsymbol{\theta}}) = (\hat{\mathbf{A}}'\hat{\mathbf{A}})^{-1}/n。$$

关于$\hat{\boldsymbol{\theta}}$的渐近正态性及其协方差可以更简便地从最大似然估计值的一般结果中得出。但是，该方法要求比这里的假定更强的规则性条件（Rao，1973，p. 364）。假如观测值来自某个独立的概率密度函数$f(\mathbf{y};\boldsymbol{\theta})$，那么，最大似然估计值$\hat{\boldsymbol{\theta}}$是有效的，也即

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\xrightarrow{d}N(\mathbf{0},\mathfrak{I}^{-1}),$$

其中$\mathfrak{I}$是单一观测值对应的信息矩阵。它的第(j,k)个元素为

$$-E\left(\frac{\partial^2\log f(\mathbf{y},\boldsymbol{\theta})}{\partial\theta_j\theta_k}\right) = E\left(\frac{\partial\log f(\mathbf{y},\boldsymbol{\theta})}{\partial\theta_j}\cdot\frac{\partial\log f(\mathbf{y},\boldsymbol{\theta})}{\partial\theta_k}\right)。$$

当f是服从多项分布概率$\{\pi_1(\boldsymbol{\theta}),\cdots,\pi_N(\boldsymbol{\theta})\}$的单一观测值的概率时，$\mathfrak{I}$的相应元素等于

$$\sum_{i=1}^N\frac{\partial\log(\pi_i(\theta))}{\partial\theta_j}\frac{\partial\log(\pi_i(\boldsymbol{\theta}))}{\partial\theta_k}\pi_i(\boldsymbol{\theta}) = \sum_{i=1}^N\frac{\partial\pi_i(\boldsymbol{\theta})}{\partial\theta_j}\frac{\partial\pi_i(\boldsymbol{\theta})}{\partial\theta_k}\frac{1}{\pi_i(\boldsymbol{\theta})}。$$

这也就是$\mathbf{A}'\mathbf{A}$的第(j,k)个元素，因而渐近协方差矩阵为$\mathfrak{I}^{-1}=(\mathbf{A}'\mathbf{A})^{-1}$。

在运用本节的结果时，$\boldsymbol{\theta}$的最大似然估计值必须存在，而且它必须是似然方程的解。这要求满足以下的*强可识别性*（*strong identifiability*）条件：对于任意$\varepsilon>0$，存在$\delta>0$，以使得如果$\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\varepsilon$，那么$\|\boldsymbol{\pi}(\boldsymbol{\theta})-\boldsymbol{\pi}_0\|>\delta$。这个条件隐含着一个较弱的条件，即两个不同的$\boldsymbol{\theta}$值不能具有相同的$\boldsymbol{\pi}$值。当强可识别性条件和其他规则性条件成立时，随着$n\to\infty$，最大似然估计值是似然方程的解的概率收敛于1。该估计值具有上文所提到的有关似然方程的解的所有渐近特性。有关证明，参见：Birch（1964a）、Rao（1973，pp. 360－362）。

14.2.2　单元格概率估计值的渐近分布

模型估计值$\hat{\boldsymbol{\pi}}$的渐近分布来自于泰勒级数展开

$$\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\pi}(\boldsymbol{\theta}_0) + \frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + o_p(n^{-1/2})。\tag{14.17}$$

剩余项的大小由$(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=O_p(n^{-1/2})$推出。现在，$\boldsymbol{\pi}(\boldsymbol{\theta}_0)=\boldsymbol{\pi}_0$，并且$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$渐近服从协方差矩阵为$(\mathbf{A}'\mathbf{A})^{-1}$的正态分布。根据$\delta$方法，

$$\sqrt{n}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0)\xrightarrow{d}N\left[\mathbf{0},\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\theta}_0}(\mathbf{A}'\mathbf{A})^{-1}\frac{\partial\boldsymbol{\pi}'}{\partial\boldsymbol{\theta}_0}\right]。\tag{14.18}$$

当模型成立且$\boldsymbol{\theta}$具有$q<N-1$个元素时，用$\hat{\boldsymbol{\pi}}=\boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$来估计$\boldsymbol{\pi}$比使用样本比例$\mathbf{p}$更有效。更一般地说，在估计关于$\boldsymbol{\pi}$的平滑函数$g(\boldsymbol{\pi})$时，$g(\hat{\boldsymbol{\pi}})$比$g(\mathbf{p})$的渐近方差更小。接下来，对这个曾在第6.4.5节提及的结果，我们进行推导。在推导过程中，我们删除了

$\mathbf{p}$ 和 $\hat{\boldsymbol{\pi}}$ 的第 N 项，以使得它们的协方差矩阵为正定矩阵（习题 14.16）。第 N 个比例线性依赖于前 $N-1$ 个比例，因为它们相加等于 1。令 $\boldsymbol{\Sigma}=\mathbf{diag}(\boldsymbol{\pi})-\boldsymbol{\pi\pi}'$ 表示 $\sqrt{n}\,\mathbf{p}$ 的 $(N-1)\times(N-1)$ 协方差矩阵。$\boldsymbol{\Sigma}$ 的逆矩阵为

$$\boldsymbol{\Sigma}^{-1}=\mathbf{diag}(\boldsymbol{\pi})^{-1}+\mathbf{11}'\pi_N, \qquad (14.19)$$

这可以通过推导 $\boldsymbol{\Sigma\Sigma}^{-1}$ 等于单位矩阵来加以验证。

令 $(\partial g/\partial\boldsymbol{\pi}_0)=(\partial g/\partial\pi_1,\cdots,\partial g/\partial\pi_{N-1})'$，求其在 $\boldsymbol{\pi}=\boldsymbol{\pi}_0$ 时的取值。根据 δ 方法，

$$\text{asymp. var}[\sqrt{n}g(\mathbf{p})]=\left(\frac{\partial g}{\partial\boldsymbol{\pi}_0}\right)'[\text{cov}(\sqrt{n}\mathbf{p})]\frac{\partial g}{\partial\boldsymbol{\pi}_0}=\left(\frac{\partial g}{\partial\boldsymbol{\pi}_0}\right)'\boldsymbol{\Sigma}\frac{\partial g}{\partial\boldsymbol{\pi}_0},$$

并且

$$\begin{aligned}\text{Asymp. var}[\sqrt{n}g(\hat{\boldsymbol{\pi}})]&=\left(\frac{\partial g}{\partial\boldsymbol{\pi}_0}\right)'[\text{Asymp. cov}(\sqrt{n}\hat{\boldsymbol{\pi}})]\frac{\partial g}{\partial\boldsymbol{\pi}_0}\\&=\left(\frac{\partial g}{\partial\boldsymbol{\pi}_0}\right)'\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\theta}_0}[\text{Asymp. cov}(\sqrt{n}\hat{\boldsymbol{\theta}})]\left(\frac{\partial\pi}{\partial\boldsymbol{\theta}_0}\right)\frac{\partial g}{\partial\boldsymbol{\pi}_0}。\end{aligned}$$

由式 14.11 和式 14.19 得出：

$$\begin{aligned}\text{Asymp. cov}(\sqrt{n}\hat{\boldsymbol{\theta}})&=(\mathbf{A}'\mathbf{A})^{-1}=[(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1}(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)]^{-1}\\&=[(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)'\boldsymbol{\Sigma}^{-1}(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)]^{-1}。\end{aligned}$$

因为 $\boldsymbol{\Sigma}$ 是正定的，并且 $(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)$ 的秩等于 q，$\boldsymbol{\Sigma}^{-1}$ 和 $[(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)'\boldsymbol{\Sigma}^{-1}(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)]^{-1}$ 也都是正定的。

为了表明 $\text{asymp. var}[\sqrt{n}g(\mathbf{p})]\geqslant\text{asymp. var}[\sqrt{n}g(\hat{\boldsymbol{\pi}})]$，我们证明

$$\left(\frac{\partial g}{\partial\boldsymbol{\pi}_0}\right)'\left\{\boldsymbol{\Sigma}-\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\theta}_0}\left[\left(\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\theta}_0}\right)'\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\theta}_0}\right]^{-1}\left(\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\theta}_0}\right)'\right\}\frac{\partial g}{\partial\boldsymbol{\pi}_0}\geqslant 0。$$

但是这个二次项形式等同于

$$(\mathbf{Y}-\mathbf{B}\xi)'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{B}\xi),$$

其中 $\mathbf{Y}=\boldsymbol{\Sigma}(\partial g/\partial\boldsymbol{\pi}_0)$，$\mathbf{B}=(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)$，$\xi=(\mathbf{B}'\boldsymbol{\Sigma}^{-1}\mathbf{B})^{-1}\mathbf{B}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$。这时，由 $\boldsymbol{\Sigma}^{-1}$ 的正定性可以推出相应结果。

这个推导过程基于 Altham(1984)所给出的证明。她在证明中应用了最大似然估计值的标准特性。只要保证这些特性成立的规则性条件都满足，这个结果就成立。该结果不仅适用于分类数据，它还适用于任意描述一组参数 $\boldsymbol{\pi}$ 对另一组更少参数 $\boldsymbol{\theta}$ 的依赖情况的模型。

14.3 残差和拟合优度统计量的渐近分布

接下来，我们讨论多项分布模型 $\boldsymbol{\pi}=\boldsymbol{\pi}(\boldsymbol{\theta})$ 的拟合优度统计量——皮尔逊 X^2 以及似然比 G^2 的分布。首先，我们推导样本比例 $\mathbf{p}$ 和模型估计值 $\hat{\boldsymbol{\pi}}$ 的渐近联合分布。这个分布决定了与 $\mathbf{p}$ 和 $\hat{\boldsymbol{\pi}}$ 有关的统计量的大样本分布。例如，它决定了皮尔逊残差的渐近联合分布，该残差用来比较 $\mathbf{p}$ 与 $\hat{\boldsymbol{\pi}}$ 的差异。这时，可以很容易地推导 X^2（即皮尔逊残差的平方和）的大样本卡方分布。另外，我们还证明，当模型成立时，X^2 和 G^2 是渐近等价的。本节的内容参考了：Bishop et al.(1975，Chap. 14)、Cox(1984)、Cramér(1946，pp. 432－433)、Rao(1973，Sect. 6b)。

14.3.1 $\mathbf{p}$ 和 $\hat{\boldsymbol{\pi}}$ 的联合渐近正态性

为了证明 $\mathbf{p}$ 和 $\hat{\boldsymbol{\pi}}$ 的联合渐近正态性，我们首先表达出 $\mathbf{p}$ 和 $\hat{\boldsymbol{\pi}}$ 对 $\mathbf{p}$ 的联合依赖性。令

$$\mathbf{D} = \mathbf{diag}(\boldsymbol{\pi}_0)^{1/2}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}。$$

由式 14.15 和式 14.17 可得，

$$\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0 = \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathbf{o}_p(n^{-1/2})$$

$$= \mathbf{D}(\mathbf{p} - \boldsymbol{\pi}_0) + \mathbf{o}_p(n^{-1/2})。$$

因此，

$$\sqrt{n}\begin{pmatrix}\mathbf{p} - \boldsymbol{\pi}_0 \\ \hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0\end{pmatrix} = \begin{pmatrix}\mathbf{I} \\ \mathbf{D}\end{pmatrix}\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}_0) + o_p(1),$$

其中 $\mathbf{I}$ 是一个 $N \times N$ 单位矩阵。根据 δ 方法，

$$\sqrt{n}\begin{pmatrix}\mathbf{p} - \boldsymbol{\pi}_0 \\ \hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0\end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}^*), \tag{14.20}$$

其中

$$\boldsymbol{\Sigma}^* = \begin{pmatrix}\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}'_0 & [\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}'_0]\mathbf{D}' \\ \mathbf{D}[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}'_0] & \mathbf{D}[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}'_0]\mathbf{D}'\end{pmatrix}。 \tag{14.21}$$

在 $\boldsymbol{\Sigma}^*$ 的主对角线上，两个子矩阵分别为上文中推导的 $\mathrm{cov}(\sqrt{n}\mathbf{p})$ 和 asymp. $\mathrm{cov}(\sqrt{n}\hat{\boldsymbol{\pi}})$。这里新的信息是 asymp. $\mathrm{cov}(\sqrt{n}\mathbf{p}, \sqrt{n}\hat{\boldsymbol{\pi}}) = [\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}'_0]\mathbf{D}'$。

14.3.2 皮尔逊和标准化残差的渐近分布

单元格计数 $\{n_i\}$ 的皮尔逊统计量为 $X^2 = \sum e_i^2$，其中

$$e_i = \frac{n_i - \hat{\mu}_i}{\hat{\mu}_i^{1/2}} = \frac{\sqrt{n}(p_i - \hat{\pi}_i)}{\hat{\pi}_i^{1/2}}。$$

接下来，我们推导 $\mathbf{e} = (e_1, \cdots, e_N)'$ 的渐近分布，用以诊断拟合不充分的情况。在泊松模型中，$\mathbf{e} = (e_1, \cdots, e_N)'$ 为皮尔逊残差，将其除以标准误就得到了相应的标准化残差。$\mathbf{e}$ 的分布也有助于推导 X^2 的分布。

残差 $\mathbf{e}$ 是 $\mathbf{p}$ 和 $\hat{\boldsymbol{\pi}}$ 的函数，根据式 14.20，二者具有联合渐近正态性。在应用 δ 方法时，我们计算

$$\partial e_i / \partial p_i = \sqrt{n}\hat{\pi}_i^{-1/2}, \qquad \partial e_i / \partial \hat{\pi}_i = -\sqrt{n}(p_i + \hat{\pi}_i)/2\hat{\pi}_i^{-3/2},$$

$$当 i \neq j 时, \qquad \partial e_i / \partial p_j = \partial e_i / \partial \hat{\pi}_j = 0。$$

也即，

$$\frac{\partial \mathbf{e}}{\partial \mathbf{p}} = \sqrt{n}\mathbf{diag}(\hat{\boldsymbol{\pi}})^{-1/2} 以及$$

$$\frac{\partial \mathbf{e}}{\partial \hat{\boldsymbol{\pi}}} = -\left(\frac{1}{2}\right)\sqrt{n}[\mathbf{diag}(\mathbf{p}) + \mathbf{diag}(\hat{\boldsymbol{\pi}})]\mathbf{diag}(\hat{\boldsymbol{\pi}})^{-3/2}。 \tag{14.22}$$

在 $\mathbf{p} = \boldsymbol{\pi}_0$ 和 $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}_0$ 处求这些偏导数的取值，相应的矩阵分别等于 $\sqrt{n}\,\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}$ 和 $-\sqrt{n}\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}$。利用式 14.21，式 14.22，以及 $\mathbf{A}'\boldsymbol{\pi}_0^{1/2} = 0$（由式 14.11 推出），$\delta$ 方法

表明

$$\mathbf{e} \xrightarrow{d} N(\mathbf{0}, \mathbf{I} - \boldsymbol{\pi}_0^{1/2}\boldsymbol{\pi}_0^{1/2'} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')。\quad (14.23)$$

极限分布的形式为 $N(\mathbf{0}, \mathbf{I} - \mathbf{Hat})$，其中 **Hat** 代表*帽子矩阵*（*hat matrix*）（第 4.5.5 节）。尽管 **e** 具有渐近正态性，它的方差却小于标准正态随机变量的方差。将 **e** 除以其标准误的估计值，就得到了标准化皮尔逊残差（Haberman, 1973a）。这个统计量渐近服从标准正态分布，可表示为

$$r_i = \frac{e_i}{[1 - \hat{\pi}_i - \sum_j \sum_k (1/\hat{\pi}_i)(\partial \pi_i / \partial \hat{\theta}_j)(\partial \pi_i / \partial \hat{\theta}_k)\hat{v}^{jk}]^{1/2}}, \quad (14.24)$$

其中，$\hat{v}^{jk}$ 代表在 $(\hat{\mathbf{A}}'\hat{\mathbf{A}})^{-1}$ 中第 j 行、第 k 列所对应的元素。r_i 的分母等于 $\sqrt{1 - \hat{h}_i}$，这里第 i 个观测值的杠杆力（leverage）$\hat{h}_i$ 是关于帽子矩阵中的第 i 个对角线元素的估计。当在二维表格中进行独立性检验时，r_i 简化为式 3.13。

14.3.3 皮尔逊统计量的渐近分布

利用正态分布与卡方分布之间的以下关系，可以证明皮尔逊 X^2 统计量渐近服从卡方分布（Rao, 1973, p.188）：

令 **X** 表示均值为 **v**、协方差矩阵为 **B** 的多元正态变量，则 $(X - \mathbf{v})'C(X - \mathbf{v})$ 服从卡方分布的充要条件是 **BCBCB = BCB**，其自由度等于 **CB** 的秩。
当 **B** 为非奇异矩阵时，这个条件简化为 **CBC = C**。

皮尔逊统计量与 **e** 之间的关系为 $X^2 = \mathbf{e}'\mathbf{e}$，因而我们通过设定 **X** 为 **e**，$\mathbf{v} = \mathbf{0}$，$\mathbf{C} = \mathbf{I}$，以及 $\mathbf{B} = \mathbf{I} - \boldsymbol{\pi}_0^{1/2}\boldsymbol{\pi}_0^{1/2'} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ 来应用以上结果。由于 $\mathbf{C} = \mathbf{I}$，$(\mathbf{X} - \mathbf{v}')\mathbf{C}(\mathbf{X} - \mathbf{v}) = \{\mathbf{e}'\mathbf{e}\} = X^2$ 服从卡方分布的条件简化为 **BBB = BB**。利用 $\mathbf{A}'\boldsymbol{\pi}_0^{1/2} = \mathbf{0}$，直接计算的结果表明 **B** 是等幂的（idempotent），所以上述条件成立。由于 **e** 渐近服从多元正态分布，所以 X^2 渐近服从卡方分布。

在对称的幂等矩阵（idempotent matrices）中，矩阵的秩等于它的迹（trace）。**I** 的迹等于 N；$\boldsymbol{\pi}_0^{1/2}\boldsymbol{\pi}_0^{1/2'}$ 的迹等于 $\boldsymbol{\pi}_0^{1/2'}\boldsymbol{\pi}_0^{1/2}$ 的迹、等于 $\sum \pi_{i0} = 1$，即为 1；$\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ 的迹等于 $(\mathbf{A}'\mathbf{A})^{-1}(\mathbf{A}'\mathbf{A})$ 的迹、等于 $q \times q$ 的单位矩阵的迹，即为 q。因此，$\mathbf{B} = \mathbf{CB}$ 的秩为 $N - q - 1$，进而，渐近的卡方分布的自由度 $df = N - q - 1$。

这个结果非常简单，由 Fisher（1922）推导给出。在大样本的情况下，X^2 的分布不依赖于 $\boldsymbol{\pi}_0$ 或模型形式，它只取决于 $\boldsymbol{\pi}$ 的维度（即 $N - 1$）与 $\boldsymbol{\theta}$ 的维度之差。在 $q = 0$ 个参数的情况下，X^2 就是皮尔逊（1900）用于检验多项分布概率等于一组特定值的统计量（式 1.15），如皮尔逊所述，它的自由度 $df = N - 1$。Watson（1959）证明，给定冗余参数的充分统计量，以上结果对渐近的条件分布也成立。

14.3.4 似然比统计量的渐近分布

当模型成立时，随着 $n \to \infty$，似然比统计量 G^2 与 X^2 渐近等价。为了对此进行证明，我们利用表达式

$$G^2 = 2\sum_i n_i \log \frac{n_i}{\hat{\mu}_i} = 2n \sum_i p_i \log\left(1 + \frac{p_i - \hat{\pi}_i}{\hat{\pi}_i}\right),$$

以及泰勒级数展开式：

$$当 |x| < 1 时，\log(1 + x) = x - x^2/2 + x^3/3 - \cdots。$$

设定 x 为$(p_i-\hat{\pi}_i)/\hat{\pi}_i$，当模型成立时它依概率收敛于0。在大样本的情况下，

$$
\begin{aligned}
G^2 &= 2n\sum_i[\hat{\pi}_i+(p_i-\hat{\pi}_i)]\left[\frac{p_i-\hat{\pi}_i}{\hat{\pi}_i}-\left(\frac{1}{2}\right)\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i^2}+\cdots\right] \\
&= 2n\sum_i\left[(p_i-\hat{\pi}_i)-\left(\frac{1}{2}\right)\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i}+\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i}+O_p(p_i-\hat{\pi}_i)^3\right] \\
&= n\sum_i\frac{(p_i-\hat{\pi}_i)^2}{\hat{\pi}_i}+2nO_p(n^{-3/2}) = X^2+O_p(n^{-1/2}) = X^2+o_p(1),
\end{aligned}
$$

由于 $\sum(p_i-\hat{\pi}_i)=0$ 以及$(p_i-\hat{\pi}_i)=(p_i-\pi_i)-(\hat{\pi}_i-\pi_i)$，二者均为$O_p(n^{-1/2})$。因此，当模型成立时，$X^2$ 与 G^2 之差依概率收敛于0。这时，与 X^2 一样，G^2 也渐近服从 $\mathrm{df}=N-q-1$ 的卡方分布。

最大化似然函数的参数值同时也使 G^2 取最小值。对其进行证明，我们令

$$G^2(\boldsymbol{\pi};\mathbf{p}) = 2n\sum p_i\log(p_i/\pi_i)。$$

多项分布对数似然函数的核函数为

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= n\sum p_i\log\pi_i(\boldsymbol{\theta}) \\
&= -n\sum p_i\log\frac{p_i}{\pi_i(\boldsymbol{\theta})}+n\sum p_i\log p_i \\
&= -\left(\frac{1}{2}\right)G^2(\boldsymbol{\pi}(\boldsymbol{\theta});\mathbf{p})+n\sum p_i\log p_i。
\end{aligned}
$$

最后一个表达式中的第二项不取决于 $\boldsymbol{\theta}$，所以最大化 $L(\boldsymbol{\theta})$ 等价于针对 $\boldsymbol{\theta}$ 最小化 G^2。

关于 G^2 的基本结果常用于进行嵌套模型间的比较。假定模型 M_0 是模型 M_1 的一个特例，令 q_0 和 q_1 分别表示两个模型的参数数量，$\{\hat{\pi}_{0i}\}$ 和 $\{\hat{\pi}_{1i}\}$ 分别表示两个模型关于单元格概率的最大似然估计值，那么，

$$G^2(M_0)-G^2(M_1) = 2n\sum p_i\log(\hat{\pi}_{1i}/\hat{\pi}_{0i})$$

具有以 M_1 成立为备择假设来检验 M_0 成立的 -2(对数似然比)的形式。似然比检验的理论表明，当较简单的模型成立时，$G^2(M_0)-G^2(M_1)$ 渐近服从自由度为 q_1-q_0 的卡方分布。有关细节，参见：Bishop et al.（1975，pp. 525-526）、Haberman（1974a，p. 108）、Rao（1973，pp. 418-419）。在式9.4中定义的 $X^2(M_0|M_1)$ 是对模型的 G^2 之差进行的二次项近似。Haberman（1977a）指出，只要与样本规模相比，q_1-q_0 很小，并且期望频数之间具有相同的数量级，即便针对大型的稀疏表格，这些检验仍然适用。

14.3.5 渐近非中心分布

本章推导的结果假定某个特定的参数模型成立。在现实应用中，任何非饱和模型几乎都不可能完全成立，所以有人可能会对这些结果存有疑虑。如果我们将模型仅仅看作是对现实世界的一种方便的近似，这就不是什么问题。例如，最大似然估计值 $\hat{\boldsymbol{\theta}}$ 收敛于值 $\boldsymbol{\theta}_0$，在该值处，所选取的模型对现实世界拟合得最好。从这个意义上来说，关于 $\boldsymbol{\theta}$ 的推断给我们提供了对现实世界的一种有意义的近似。相似地，当模型不成立时，有关单元格概率的模型推断与真实的概率并不一致，不过，这些推断仍是对现实情况的一种有益修匀。

在模型成立和不成立的情况下,拟合优度统计量的极限特性存在区别。当模型成立时,我们已经看到 X^2 和 G^2 的极限服从卡方分布,并且随着 n 的上升二者之间的差异消失。当模型不成立时,随着 n 的上升,X^2 与 G^2 一般持续增大,而且 $|X^2-G^2|$ 不一定会趋近于零。获取适当极限分布的一种方法是,考虑一系列情况对应的 $\boldsymbol{\pi}_n$,其中拟合不充分的程度随着 n 的上升而下降。具体地说,模型为 $\boldsymbol{\pi}=\mathbf{f}(\boldsymbol{\theta})$,但现实情况却是

$$\boldsymbol{\pi}_n = \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\delta}/\sqrt{n}。\tag{14.25}$$

模型对总体的最优拟合结果为第 i 个概率等于 $f_i(\boldsymbol{\theta})$,但它与真实值之间仍相差 $\delta_i/\sqrt{n}$。

Mitra(Mitra,1958)表明,在这种情况下,皮尔逊 X^2 的极限服从非中心卡方分布,其自由度 $\mathrm{df}=N-q-1$,非中心参数为

$$\lambda = n\sum_{i=1}^{n}\frac{[\pi_{ni}-f_i(\boldsymbol{\theta})]^2}{f_i(\boldsymbol{\theta})}。$$

该参数具有 X^2 的形式,其中样本值 p_i 和 $\hat{\pi}_i$ 由总体值 π_{ni} 和 $f_i(\boldsymbol{\theta})$ 所替代。与此相同,似然比统计量的非中心参数具有 G^2 的形式。Haberman(1974a,pp. 109-112)证明,在一定条件下,G^2 和 X^2 的极限服从相同的分布,也即,随着 $n\to\infty$,它们的非中心参数收敛于同一个值。

式 14.25 表明,在大样本的情况下,当模型基本正确时,非中心卡方近似是有效的。在现实应用中,当 n 为一个给定的有限值时,即便在获得更多数据后,式 14.25 可能并不合理,这时我们往往仍利用式 14.25 来近似 X^2 的分布。式 14.25 还可以表示为

$$\boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\delta},\tag{14.26}$$

其中,随着 $n\to\infty$,$\boldsymbol{\pi}$ 与 $\mathbf{f}(\boldsymbol{\theta})$ 之间相差一个常数。这似乎是一种更自然的表述。事实上,从提供具有一致性检验(即,拒绝模型成立的假设的概率收敛于 1)的角度来说,式 14.26 比式 14.25 更适当。但是,在式 14.26 中,随着 $n\to\infty$,非中心参数 λ 持续增大,并且 X^2 和 G^2 不存在适当的极限分布。

当模型成立时,无论在式 14.25 还是式 14.26 中,都有 $\boldsymbol{\delta}=0$。也即,$\mathbf{f}(\boldsymbol{\theta})=\boldsymbol{\pi}(\boldsymbol{\theta})$,$\lambda=0$,因而可以使用第 14.3.3 和第 14.3.4 节的结果。

14.4　Logit/对数线性模型的渐近分布

在对数线性模型中,第 8.6 节关于 $\hat{\boldsymbol{\theta}}$ 和 $\hat{\boldsymbol{\pi}}$ 的渐近协方差矩阵的公式,相当于在第 14.2节所推导的结果的特例。与第 14.2 节的内容直接相关,我们给出多项分布模型的相应结果。然后,我们讨论这些结果与泊松对数线性模型的联系。

限定概率之和等于 1,我们将多项分布样本下的对数线性模型表示为

$$\boldsymbol{\pi} = \exp(\mathbf{X}\boldsymbol{\theta})/[\mathbf{1}'\exp(\mathbf{X}\boldsymbol{\theta})],\tag{14.27}$$

其中 $\mathbf{X}$ 为模型矩阵,并且 $\mathbf{1}'=(1,\cdots,1)$。令 $\mathbf{x}_i$ 表示 $\mathbf{X}$ 的第 i 行,那么

$$\pi_i = \pi_i(\boldsymbol{\theta}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\theta})}{\sum_k\exp(\mathbf{x}_k\boldsymbol{\theta})}。$$

14.4.1　渐近协方差矩阵

模型通过雅可比矩阵来求解协方差矩阵。由于

$$\frac{\partial\pi_i}{\partial\theta_j}=\frac{[\sum_k\exp(\mathbf{x}_k\boldsymbol{\theta})][\exp(\mathbf{x}_i\boldsymbol{\theta})]x_{ij}-[\exp(\mathbf{x}_i\boldsymbol{\theta})][\sum_k x_{kj}\exp(\mathbf{x}_k\boldsymbol{\theta})]}{[\sum_k\exp(\mathbf{x}_k\boldsymbol{\theta})]^2}$$

$$= \pi_i x_{ij} - \pi_i \sum_k x_{kj}\pi_k,$$

这些元素的矩阵形式为

$$\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta} = [\mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi\pi}']\mathbf{X}。$$

利用这个表达式以及式 14.14 和式 14.16，在 $\boldsymbol{\theta}_0$ 处的信息矩阵为

$$\begin{aligned}\mathbf{A}'\mathbf{A} &= (\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1}(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)\\ &= \mathbf{X}'[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0']'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1}[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0']\mathbf{X}\\ &= \mathbf{X}'[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0']\mathbf{X}。\end{aligned}$$

因此，在多项分布对数线性模型中，$\hat{\boldsymbol{\theta}}$ 渐近服从正态分布，对其估计的协方差矩阵为

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\theta}}) = \{\mathbf{X}'[\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}']\mathbf{X}\}^{-1}/n。 \qquad (14.28)$$

相似地，由式 14.23 可得，所估计的关于 $\hat{\boldsymbol{\pi}}$ 的渐近协方差矩阵为

$$\begin{aligned}\widehat{\mathrm{cov}}(\hat{\boldsymbol{\pi}}) &= [\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}']\mathbf{X}\{\mathbf{X}'[\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}']\mathbf{X}\}^{-1}\\ &\quad\times \mathbf{X}'[\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}']/n。\end{aligned}$$

根据式 14.23，皮尔逊残差 $\mathbf{e}$ 渐近服从正态分布，并有

$$\begin{aligned}\text{asymp. cov}(\mathbf{e}) &= \mathbf{I} - \boldsymbol{\pi}_0^{1/2}(\boldsymbol{\pi}_0^{1/2})' - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\\ &= \mathbf{I} - \boldsymbol{\pi}_0^{1/2}(\boldsymbol{\pi}_0^{1/2})' - \mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0']\mathbf{X}\\ &\quad\times\{\mathbf{X}'[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0']\mathbf{X}\}^{-1}\mathbf{X}'\\ &\quad\times[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0']\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}。\end{aligned}$$

14.4.2 与泊松对数线性模型的联系

在本书中，对数线性模型用服从泊松分布的期望单元格频数 $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_N)'$ 来表示，其公式为

$$\log\boldsymbol{\mu} = \mathbf{X}_a\boldsymbol{\theta}_a。 \qquad (14.29)$$

上述公式中的模型矩阵 $\mathbf{X}_a$ 和参数向量 $\boldsymbol{\theta}_a$ 与多项分布模型（式 14.27）中的 $\mathbf{X}$ 和 $\boldsymbol{\theta}$ 略有不同。在泊松表达式（式 14.29）中，不包括关于 $\boldsymbol{\mu}$ 的限定条件。而对于多项分布模型（式 14.27），$\sum_i \mu_i = n$ 是给定的，并且 $\boldsymbol{\pi} = \boldsymbol{\mu}/n$ 满足

$$\begin{aligned}\log\boldsymbol{\mu} = \log n\boldsymbol{\pi} &= \mathbf{X}\boldsymbol{\theta} + [\log n - \log(\mathbf{1}'\exp(\mathbf{X}\boldsymbol{\theta}))]\mathbf{1}\\ &= \mathbf{X}\boldsymbol{\theta} + \mathbf{1}\mu\end{aligned}$$

其中 $\mu = \log n - \log[\mathbf{1}'\exp(\mathbf{X}\boldsymbol{\theta})]$。换句话说，多项分布模型（式 14.27）意味着在泊松模型（式 14.29）中存在

$$\mathbf{X}_a = [\mathbf{1}:\mathbf{X}] \quad \text{并且} \quad \boldsymbol{\theta}_a = (\mu, \boldsymbol{\theta}')'。$$

在多项分布表达式中，$\mathbf{X}$ 的列必须线性独立于 $\mathbf{1}$，也即，与总的样本规模有关的参数 μ 不出现在 $\boldsymbol{\theta}$ 里，其中，$\boldsymbol{\theta}$ 的维度比本书所介绍的泊松对数线性模型的参数数量少 1 个。例如，就饱和模型而言，在多项分布表达式中 $\boldsymbol{\theta}$ 具有 $N-1$ 个元素，它反映了关于 $\boldsymbol{\pi}$ 的单一限定条件 $\sum \pi_i = 1$。

注　解

第 14.1 节：δ 方法

14.1　有关包括 δ 方法的大样本理论的详细讨论，参见：Bishop et al.（1975，chap. 14）、

Sen and Singer(1993)。

14.2 在对渐近正态随机向量 $\mathbf{T}_n$ 的函数 g 应用 δ 方法时,假定该函数的第一阶,……,第 $(a-1)$ 阶导数在 $\boldsymbol{\theta}$ 处都等于零,但第 a 阶导数不为零,对 δ 方法的扩展表明,$n^{a/2}[g(\mathbf{T}_n)-g(\boldsymbol{\theta})]$ 的极限服从与正态随机向量的第 a 阶项有关的分布。当 $a=2$ 时,极限分布是多元正态向量的一个二次项,并常与卡方分布具有一定的联系;在单变量的情况下,它等于 $\sigma^2(g''(\theta))/2$ 乘以一个 χ_1^2 变量(Casella and Berger, 2001, pp. 244)。

再抽样法,如刀切法和重抽样自举法,是估计标准误与获取置信区间的另一种方法。当 δ 方法不适用时,比如小样本的情况、数据高度稀疏或样本来自于复杂抽样设计,再抽样法非常有意义。有关细节,参见:Davison and Hinkley(1997)、Fay(1985)、Parr and Tolley(1982)、Simonoff(1986)。

第 14.3 节:残差和拟合优度统计量的渐近分布

14.3 如果 Y 是一个均值为 $E(Y)=\mu$ 的泊松变量,那么当 μ 很大时,δ 方法表明,$Y^{1/2}$ 近似服从标准差为 $\frac{1}{2}$ 的正态分布。由此可以推出另一个拟合优度统计量,即 Freeman-Tukey 统计量(*the Freeman-Tukey statistic*),$\mathrm{FT}=4\sum(\sqrt{y_i}-\sqrt{\hat{\mu}_i})^2$。当模型成立时,随着 $n\to\infty$,$\mathrm{FT}-X^2$ 也满足 $o_p(1)$。有关细节,参见:Bishop et al. (1975, p. 514)。

当单元格数量 N 随着 $n\to\infty$ 而上升或者当不同单元格期望频数按照不同的速度增加时,本章的结果不适用。Haberman(1988)表明,这种情况下 X^2 的一致性不再成立,且不具有标准的渐近特性。

14.4 与式 14.25 和式 14.26 所使用的局部固定序列不同,Drost 等(1989)给出了其他序列下的非中心近似。

习 题

14.1 解释原因:

a. 如果 $c>0$,那么随着 $n\to\infty$,$n^{-c}=o(1)$。

b. 如果 $c\neq 0$,cz_n 具有与 z_n 相同的阶数,也即,$o(cz_n)$ 等价于 $o(z_n)$,并且 $O(cz_n)$ 等价于 $O(z_n)$。

c. $o(y_n)o(z_n)=o(y_nz_n)$,$O(y_n)O(z_n)=O(y_nz_n)$,$o(y_n)O(z_n)=o(y_nz_n)$。

14.2 随着 $n\to\infty$,如果 X^2 渐近服从具有固定自由度的卡方分布,那么说明为什么 $X^2/n=o_p(1)$。

14.3 a. 利用切比雪夫不等式(Tchebychev ' s ineuqality)证明,如果 $E(X_n)=\mu_n$ 并且 $\mathrm{var}(X_n)=\sigma_n^2<\infty$,那么 $(X_n-\mu_n)=O_p(\sigma_n)$。

b. 假定对于 $i=1,\cdots,n$,$Y_1,\cdots,Y_n$ 为相互独立的变量,其中 $E(Y_i)=\mu$ 以及 $\mathrm{var}(Y_i)=\sigma^2$,令 $\bar{Y}_n=(\sum_i Y_i)/n$,利用(a)部分的结果证明 $\bar{Y}_n-\mu=O_p(n^{-1/2})$。

14.4 令 Y 表示一个均值为 μ 的泊松随机变量。

a. 对于常数 $c>0$,证明

$$E[\log(Y+c)]=\log\mu+\left(c-\frac{1}{2}\right)/\mu+O(\mu^{-2})。$$

(提示:注意 $\log(Y+c)=\log\mu+\log[1+(Y+c-\mu)/\mu]$)。

b. 在一个 2×2 表格中,单元格计数是独立的泊松随机变量。利用(a)部分的结果证明,在估计对数发生比之比时为了降低偏差,合理的估计值是将每个单元格计数分别加上$\frac{1}{2}$后的样本对数发生比之比。

14.5 令 p 代表 n 个独立伯努利试验的样本比例,求相应标准差的估计值$[p(1-p)]^{1/2}$的渐近分布。当 $\pi=0.5$ 时,会出现什么情况?

14.6 对于给定的 $\mu>0$,假设 T_n 服从均值 $\lambda=n\mu$ 的泊松分布。在大样本的情况下,证明 $\log T_n$ 近似服从均值为 $\log(\lambda)$、方差为 λ^{-1} 的正态分布(提示:按照中心极限定理,在大样本情况下 T_n/n 近似服从 $N(\mu,\mu/n)$)。

14.7 a. 参考习题 14.6。如果 T_n 服从泊松分布,证明 $\sqrt{T_n}$ 的渐近方差等于$\frac{1}{4}$。

b. 对于试验次数为 n 和样本比例为 p 的二项分布样本,证明 $\sin^{-1}(\sqrt{p})$ 的渐近方差等于 $1/4n$(这个转换与(a)部分的转换都是*方差恒定的*(*variance stabilizing*),所生成变量的渐近方差对参数的所有取值都相同。过去,这些转换曾被用来对计数数据进行普通最小二乘法估计。有关讨论以及最大似然分析,参见:Cochran (1940))。

14.8 对于多项分布$(n,\{\pi_i\})$,证明 p_i 和 p_j 之间的相关系数等于 $-[\pi_i\pi_j/(1-\pi_i)(1-\pi_j)]^{1/2}$。当 $\pi_i=1-\pi_j$ 且对 $k\neq i,j$ 有 $\pi_k=0$ 时,这个相关系数等于多少?

14.9 某一动物总体包括 N 种不同的动物,其中第 i 种在总体中所占的比例为 π_i。*辛普森生态多样性指数*(*Simpson's index of ecological diversity*)(Simpson, 1949)为 $I(\boldsymbol{\pi})=1-\sum\pi_i^2$(Rao(1982)回顾了有关的多样性测量指标)。

a. 从总体中有放回地随机选取两只动物,证明 $I(\boldsymbol{\pi})$ 等于它们来自于不同种类的概率。

b. 对于一个随机样本的样本比例 $\mathbf{p}$,证明所估计的 $I(\mathbf{p})$ 的渐近标准误为

$$2\left\{\left[\sum_i p_i^3-\left(\sum_i p_i^2\right)^2\right]/n\right\}^{1/2}。$$

14.10 令$\{Y_i\}$表示独立的泊松随机变量,利用 δ 方法证明,所估计的关于 $\sum a_i\log(Y_i)$ 的渐近方差等于 $\sum a_i^2/y_i$(这个公式适用于对饱和对数线性模型参数的最大似然估计值,它与$\{\log(y_i)\}$相对应。式 14.9 给出了这些估计值的渐近协方差结构;参见:Lee(1977))。

14.11 假定存在两个独立的二项分布样本,推导出它们的对数相对风险(第 3.1.4 节)的渐近标准误。

14.12 参考习题 3.27。为了使 $\hat{\gamma}$ 的样本分布近似服从正态分布,可能需要非常大的样本规模,尤其是当 $|\gamma|$ 很大时。相应的 Fisher 转换 $\hat{\xi}=\frac{1}{2}\log[(1+\hat{\gamma})/(1-\hat{\gamma})]$(Agresti, 1984: pp. 166-167, 177; O'Gorman and Woolson, 1988)以更快的速度收敛于正态分布。

a. 证明 $\hat{\xi}$ 的渐近方差等于 $\hat{\gamma}$ 的渐近方差的$(1-\gamma^2)^{-2}$倍。

b. 说明如何构建一个关于 ξ 的置信区间,并用此推导关于 γ 的置信区间。

c. 证明 $\hat{\xi}=\frac{1}{2}\log(C/D)$。在 2×2 表格中,证明它等于对数发生比之比的一半。

14.13 令 $\phi^2(\mathbf{T}) = \sum_i (T_i - \pi_{i0})^2/\pi_{i0}$，这时，$\phi^2(\mathbf{p}) = X^2/n$，其中 X^2 是检验 $H_0: \pi_i = \pi_{i0}, i = 1, \cdots, N$ 的皮尔逊统计量（式 1.15），并且 $n\phi^2(\boldsymbol{\pi})$ 是当 $\boldsymbol{\pi}$ 等于真实值时该检验的非中心参数。在 H_0 下，为什么 δ 方法无法推出 $\phi^2(\mathbf{p})$ 渐近服从正态分布？（参见注解 14.2）

14.14 在一个 $I \times J$ 列联表中，令 θ_{ij} 代表局部发生比之比（式 2.10），$\hat{\theta}_{ij}$ 表示其样本值。

a. 证明 asymp. cov$(\sqrt{n}\log\hat{\theta}_{ij}, \sqrt{n}\log\hat{\theta}_{i+1,j}) = -[\pi_{i+1,j}^{-1} + \pi_{i+1,j+1}^{-1}]$。

b. 证明 asymp. cov$(\sqrt{n}\log\hat{\theta}_{ij}, \sqrt{n}\log\hat{\theta}_{i+1,j+1}) = \pi_{i+1,j+1}^{-1}$。

c. 当 $\hat{\theta}_{ij}$ 和 $\hat{\theta}_{hk}$ 所使用的单元格不重叠时，证明 asymp. cov$(\sqrt{n}\log\hat{\theta}_{ij}, \sqrt{n}\log\hat{\theta}_{hk}) = 0$。

d. 给出 $\log\hat{\theta}_{ij}$ 的渐近分布。

14.15 在对数线性模型 (XY, XZ, YZ) 中，关于 $\{\mu_{ijk}\}$ 的最大似然估计没有直接解，相应的 X^2 和 G^2 统计量也是如此。运用其他方法有可能进行直接的分析。对于 $2 \times 2 \times 2$ 表格，利用关于 $\log\hat{\theta}_{111}$ 具有渐近正态性的 δ 方法，求一个对不存在三维交互项进行假设检验的统计量，其中

$$\hat{\theta}_{111} = \frac{p_{111}p_{221}/p_{121}p_{211}}{p_{112}p_{222}/p_{122}p_{212}}。$$

14.16 参考第 14.2.2 节，$\sqrt{n}(p_1, \cdots, p_{N-1})'$ 的协方差矩阵为 $\boldsymbol{\Sigma} = \mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$。令

$$Z = \begin{cases} c_i & （概率为\ \pi_i, \quad i = 1, \cdots, N-1） \\ 0 & （概率为\ \pi_N） \end{cases},$$

并令 $\mathbf{c} = (c_1, \cdots, c_{N-1})'$。

a. 证明 $E(Z) = \mathbf{c}'\boldsymbol{\pi}$，$E(Z^2) = \mathbf{c}'\mathbf{diag}(\boldsymbol{\pi})\mathbf{c}$，以及 $\text{var}(Z) = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$。

b. 假定至少存在一个 $c_i \neq 0$，且所有 $\pi_i > 0$，证明 $\text{var}(Z) > 0$，并推导出 $\boldsymbol{\Sigma}$ 是正定的。

c. 如果 $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_N)'$，$\boldsymbol{\Sigma}$ 是 $N \times N$ 矩阵，证明 $\boldsymbol{\Sigma}$ 是非正定的。

14.17 考虑关于 2×2 表格的模型，$\pi_{11} = \theta^2$，$\pi_{12} = \pi_{21} = \theta(1-\theta)$，$\pi_{22} = (1-\theta)^2$，其中 θ 是未知的（习题 3.31 和 10.34）。

a. 求这个模型在式 14.14 中的矩阵 $\mathbf{A}$。

b. 利用 $\mathbf{A}$ 来求解 $\hat{\theta}$ 的渐近方差（为了验证这一结果，可以很容易地直接对 $-E\partial^2 L/\partial\theta^2$ 求逆来得出这个渐近方差，其中 L 为对数似然函数）。当 θ 取什么值时，方差会取最大值？如果 $\theta = 0$ 或 $\theta = 1$，$\hat{\theta}$ 的分布分别是什么？

c. 求 $\sqrt{n}\hat{\boldsymbol{\pi}}$ 的渐近协方差矩阵。

d. 给出利用 X^2 进行拟合优度检验的自由度。

14.18 参考第 1.5.6 节中有关小牛犊数据的模型，求 $\hat{\pi}$ 的渐近方差。

14.19 给出使用*估计的*（*estimated*）渐近协方差矩阵的理由。例如，在大样本的情况下，为什么 $\hat{\mathbf{A}}'\hat{\mathbf{A}}$ 近似于 $\mathbf{A}'\mathbf{A}$？

14.20 单元格计数 $\{Y_i\}$ 是独立的泊松随机变量，并有 $\mu_i = E(Y_i)$。考虑泊松对数线性模型

$$\log \boldsymbol{\mu} = \mathbf{X_a}\boldsymbol{\theta}_\mathbf{a}，其中\ \boldsymbol{\mu} = (\mu_1, \cdots, \mu_N),$$

利用第 14.2 节中的有关论断，证明可以由 $[\mathbf{X}_\mathbf{a}'\mathbf{diag}(\hat{\boldsymbol{\mu}})\mathbf{X_a}]^{-1}$ 来估计 $\hat{\boldsymbol{\theta}}_a$ 的大样本

协方差矩阵，其中 $\hat{\boldsymbol{\mu}}$ 是对 $\boldsymbol{\mu}$ 的最大似然估计值。

14.21 对于一组给定的参数限定条件，证明二维表格的独立性对数线性模型满足弱可识别性条件(weak identifiability conditions)，也即，当两个 $\boldsymbol{\theta}$ 值都给出相同的 $\boldsymbol{\pi}$ 时，这些参数向量也必定相同。

14.22 根据式 14.22 的结果，通过 δ 方法推导关于残差的渐近协方差矩阵式(14.23)。证明这个矩阵是等幂的(idempotent)。

14.23 在一定情况下，X^2 和 G^2 的值非常相似。分别说明下列条件对该结论的影响：(a)模型是否成立，(b)样本规模 n 是否很大，(c)单元格数量 N 是否很大。

14.24 按照多项分布表述形式(式 14.27)，给出关于 $I\times J$ 表格的独立性模型所对应的 $\mathbf{X}$ 和 $\boldsymbol{\theta}$。与之相对照，给出相应的泊松对数线性模型(式 14.29)中的 $\mathbf{X}_a$。

14.25 利用式 14.18 和式 14.28，推导多项分布对数线性模型的渐近$\widehat{\text{cov}}(\hat{\boldsymbol{\pi}})$。

14.26 在独立性模型中，考虑当模型不成立时关于 π_{ij} 的最大似然估计值 $\hat{\pi}_{ij}=p_{i+}p_{+j}$。证明 $E(p_{i+}p_{+j})=\pi_{i+}\pi_{+j}(n-1)/n+\pi_{ij}/n$。随着 n 的上升，$\hat{\pi}_{ij}$ 会收敛于什么值?

14.27 令 ζ 表示一般性的关联测量指标，对于样本规模为 $\{n_k\}$ 的 K 个独立的多项分布样本，假定随着 $n_k\rightarrow\infty$，$\sqrt{n_k}(\hat{\zeta}_k-\zeta_k)\xrightarrow{d}N(0,\sigma_k^2)$，相应的综合测量指标为

$$\bar{\zeta}=\frac{\sum_k(n_k/\hat{\sigma}_k^2)\hat{\zeta}_k}{\sum_k(n_k/\hat{\sigma}_k^2)}。$$

a. 证明 $\sum_k z_k^2=V+[\bar{\zeta}^2/\hat{\sigma}^2(\bar{\zeta})]$，其中

$$V=\sum_k\frac{n_k(\hat{\zeta}_k-\bar{\zeta})^2}{\hat{\sigma}_k^2},\ z_k=\frac{n_k^{1/2}\hat{\zeta}_k}{\hat{\sigma}_k},\ \hat{\sigma}^2(\bar{\zeta})=\left(\sum_k\frac{n_k}{\hat{\sigma}_k^2}\right)^{-1}。$$

b. 假定 $n\rightarrow\infty$ 且 $n_k/n\rightarrow\rho_k>0,k=1,\cdots,K$，给出在(a)部分的分割中每一项所渐近服从的卡方分布，并指出各项所检验的假设。

15 参数模型的其他估计理论

在本书中,我们主要利用了最大似然(ML)法来进行统计推断。到目前为止,最大似然法仍是分类数据分析中最常用的方法。此外,其他估计分类数据方法也有一定的应用。在本章中,我们介绍其中一些方法,这些方法具有与最大似然法相似的渐近特性,所以第14章所介绍的大样本理论对它们同样适用。

在第15.1节,我们讨论拟合分类数据模型的加权最小二乘法。在一定情况下,加权最小二乘法与第4.7节和第11.4节介绍的类似然法比最大似然法更为简便。

随着运算能力的提高,贝叶斯理论正变得越来越普遍。对贝叶斯理论发展的全面讨论超出了本书的范围,但在第15.2节中我们将介绍估计列联表单元格概率的贝叶斯方法。本章的最后一节简单介绍其他四种估计分类数据的方法。

15.1 关于分类数据的加权最小二乘法

加权最小二乘法(weighted least squares, WLS)是对普通最小二乘法的一种扩展,它允许结果变量之间存在相关关系并具有不等的方差。熟悉加权最小二乘法非常重要,这是因为:

1. 加权最小二乘法具有标准形式的运算,对许多模型而言,它的应用都很简单。
2. 计算最大似然估计的算法常常包括对加权最小二乘法的迭代使用。关于广义线性模型的Fisher计分法(第4.6.3节)就是其中的一个例子。
3. 当模型成立时,加权最小二乘法与最大似然法的估计值渐近等价,二者都属于最优渐近正态(best asymptotically normal, BAN)估计值。在大样本的情况下,估计值在参数值的两侧近似服从正态分布,且二者的方差之比收敛于1。

Grizzle, Starmer 和 Koch(1969)在分类数据分析中推广了加权最小二乘法。为了纪念他们,这类分析中的加权最小二乘法也被称为*GSK*方法(*GSK method*)。本节扼要介绍一下该方法的基本原理。

15.1.1 加权最小二乘法的标示符号与基本原理

对于具有J个类别的结果变量Y,考虑在某一解释变量或几个解释变量所组合的I个取值水平上,分别对应着规模为$n_1, \cdots, n_I$的多项分布样本。令$\boldsymbol{\pi} = (\boldsymbol{\pi}_1', \cdots, \boldsymbol{\pi}_I')'$,其中

$$\boldsymbol{\pi}_i = (\pi_{1|i}, \pi_{2|i}, \cdots, \pi_{J|i})'$$

表示在解释变量取值为 i 时 Y 的条件分布,并有 $\sum_j \pi_{j|i} = 1$。令 $\mathbf{p}$ 表示相应的样本比例,$\mathbf{V}$ 是它们之间的$IJ \times IJ$ 协方差矩阵。当这 I 个样本相互独立时,

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & & & \mathbf{0} \\ & \mathbf{V}_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{V}_I \end{bmatrix}。$$

由第 14.1.4 节可知,$\sqrt{n_i}\mathbf{p}_i$ 的协方差矩阵等于

$$n_i\mathbf{V}_i = \begin{bmatrix} \pi_{1|i}(1-\pi_{1|i}) & -\pi_{1|i}\pi_{2|i} & \cdots & -\pi_{1|i}\pi_{J|i} \\ -\pi_{2|i}\pi_{1|i} & \pi_{2|i}(1-\pi_{2|i}) & \cdots & -\pi_{2|i}\pi_{J|i} \\ \vdots & \vdots & & \vdots \\ -\pi_{J|i}\pi_{1|i} & -\pi_{J|i}\pi_{2|i} & \cdots & \pi_{J|i}(1-\pi_{J|i}) \end{bmatrix}。$$

每一组比例都具有$(J-1)$个线性独立的元素。

令 $\mathbf{F}$ 表示由 $u \leqslant I(J-1)$ 个结果变量的函数所组成的向量,

$$\mathbf{F}(\boldsymbol{\pi}) = [F_1(\boldsymbol{\pi}),\cdots,F_u(\boldsymbol{\pi})]'。$$

加权最小二乘法适用于 $\mathbf{F}$ 具有以下形式的线性模型:

$$\mathbf{F}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}, \tag{15.1}$$

其中 $\boldsymbol{\beta}$ 是一个 $q\times 1$ 的参数向量,$\mathbf{X}$ 是关于已知常数的 $u\times q$ 模型矩阵,它的秩等于 q。根据第 8.5.4 节,对于特定的矩阵 $\mathbf{C}$ 和 $\mathbf{A}$,结果变量的对数线性函数和 logit 函数是 $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{C}\log(\mathbf{A}\boldsymbol{\pi})$的特例。

令 $\mathbf{F}(\mathbf{p})$表示样本的结果变量函数。我们假定 $\mathbf{F}$ 在一个包含 $\boldsymbol{\pi}$ 的开放区域内存在连续的二阶偏导数,该假定确保可以利用 δ 方法来推导 $\mathbf{F}(\mathbf{p})$的大样本正态分布。对于 $k = 1,\cdots,u$ 以及所有的 IJ 组合(i,j),$\mathbf{F}(\mathbf{p})$的渐近协方差矩阵取决于 $u\times IJ$ 矩阵

$$\mathbf{Q} = \frac{\partial F_k(\boldsymbol{\pi})}{\partial \pi_{j|i}}。$$

线性模型中结果变量函数具有 $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{A}\boldsymbol{\pi}$ 的形式,其中 $\mathbf{A}$ 是关于已知常数的矩阵,这时 $\mathbf{Q} = \mathbf{A}$。对于广义线性模型 $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{C}\log(\mathbf{A}\boldsymbol{\pi})$(回顾第 8.5.4 节和第 11.2.5 节),$\mathbf{Q} = \mathbf{C}[\operatorname{diag}(\mathbf{A}\boldsymbol{\pi})^{-1}]\mathbf{A}$(有关矩阵的微分计算,参见:Magnus and Neudecker(1988))。根据多元的 δ 方法(第 14.1.5 节),$\mathbf{F}(\mathbf{p})$的渐近协方差矩阵等于

$$\mathbf{V}_F = \mathbf{QVQ}'。$$

令 $\hat{\mathbf{V}}_F$ 表示 $\mathbf{V}_F$ 对应的样本形式,即在 $\mathbf{Q}$ 和 $\mathbf{V}$ 中代入相应的样本比例。在下面的公式中,矩阵 $\hat{\mathbf{V}}_F$ 必须是非奇异的。

15.1.2 利用加权最小二乘法推断模型的拟合情况

在式 15.1 一般性模型中,关于 $\boldsymbol{\beta}$ 的加权最小二乘法估计为

$$\mathbf{b} = (\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{F}(\mathbf{p})。$$

这是使得二次项

$$[\mathbf{F}(\mathbf{p}) - \mathbf{X}\boldsymbol{\beta}]'\hat{\mathbf{V}}_F^{-1}[\mathbf{F}(\mathbf{p}) - \mathbf{X}\boldsymbol{\beta}]$$

取最小值的 $\boldsymbol{\beta}$ 值。当 $\hat{\mathbf{V}}_F$ 是单位矩阵的常数倍时,就得到了关于相互独立且方差恒定的结果变量的普通最小二乘法估计。

加权最小二乘法估计值渐近服从多元正态分布,相应的协方差矩阵估计值为

$$\hat{\text{cov}}(\mathbf{b}) = (\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}。$$

随着样本规模的增加以及 $\mathbf{F}(\mathbf{p})$ 更接近于正态分布，估计值的渐近正态性也会提高。

根据估计值 $\mathbf{b}$ 可以得出结果变量函数的预测值 $\hat{\mathbf{F}} = \mathbf{Xb}$。由于这些预测值满足模型条件，它们比样本对应的结果变量函数 $\mathbf{F}(\mathbf{p})$ 更平滑。当模型成立时，$\mathbf{F}(\boldsymbol{\pi})$ 的估计值 $\hat{\mathbf{F}}$ 渐近优于 $\mathbf{F}(\mathbf{p})$（第 14.2.2 节）。所估计的预测值的协方差矩阵为

$$\hat{\mathbf{V}}_{\hat{F}} = \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}\mathbf{X}'。$$

通过残差项

$$W = [\mathbf{F}(\mathbf{p}) - \mathbf{Xb}]'\hat{\mathbf{V}}_F^{-1}[\mathbf{F}(\mathbf{p}) - \mathbf{Xb}] = \mathbf{F}(\mathbf{p})\hat{\mathbf{V}}_F^{-1}\mathbf{F}(\mathbf{p}) - \mathbf{b}'(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})\mathbf{b}$$

可以对模型进行拟合优度检验，该残差项将样本的结果变量函数与其预测值进行对比。在 H_0：$\mathbf{F}(\boldsymbol{\pi}) - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ 模型成立时，W 渐近服从卡方分布，自由度为 $df = u - q$，即结果变量的函数个数与模型参数数量之差。

我们可以通过分析残差 $\mathbf{F}(\mathbf{p}) - \hat{\mathbf{F}}$ 来进一步检查模型的拟合情况。这些残差与拟合值 $\hat{\mathbf{F}}$ 相互独立，所以

$$\text{cov}[\mathbf{F}(\mathbf{p})] = \text{cov}\{[\mathbf{F}(\mathbf{p}) - \hat{\mathbf{F}}] + \hat{\mathbf{F}}\} = \text{cov}[\mathbf{F}(\mathbf{p}) - \hat{\mathbf{F}}] + \text{cov}(\hat{\mathbf{F}})。$$

这时，所估计的残差的协方差矩阵等于

$$\text{cov}[\mathbf{F}(\mathbf{p})] - \text{cov}(\hat{\mathbf{F}}) = \hat{\mathbf{V}}_F - \hat{\mathbf{V}}_{\hat{F}} = \hat{\mathbf{V}}_F - \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}\mathbf{X}'。$$

将残差除以它们的标准误就得到了标准化残差，后者服从大样本标准正态分布。

有关解释变量效应的假设具有 H_0：$\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ 的形式，其中 $\mathbf{C}$ 是一个 $c \leq q$ 的已知 $c \times q$ 矩阵，该矩阵的秩为 c。在 H_0 成立的情况下，$\mathbf{C}\boldsymbol{\beta}$ 的估计值 $\mathbf{Cb}$ 渐近服从均值为 $\mathbf{0}$ 的正态分布，所估计的协方差矩阵为 $\mathbf{C}(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}\mathbf{C}'$。沃尔德统计量

$$W_C = \mathbf{b}'\mathbf{C}'[\mathbf{C}(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})\mathbf{C}']^{-1}\mathbf{Cb} \tag{15.2}$$

近似服从 $df = c$ 的卡方零分布。这个统计量也等于由 H_0 所指定的简约模型和完整模型之间的残差卡方统计量之差。在 H_0：$\beta_i = 0$ 的特例中，$W_C = b_i^2/\text{var}(b_i)$，自由度为 $df = 1$。

15.1.3　加权最小二乘法与最大似然估计的适用范围

在解释变量的每个取值水平上，加权最小二乘法需要估计样本结果变量的多项分布协方差矩阵。当解释变量为连续变量时，这一方法不再适用，因为在每个取值处可能只有一个观测值。随着解释变量数量的增加，加权最小二乘法的适用性也会变差，因为在每个取值组合处的观测值数量越来越少。与之相反，原则上最大似然法完全可以处理连续的解释变量或解释变量包括许多不同取值的情况。

在单元格期望频数很大的情况下，当某个模型成立时，最大似然法和加权最小二乘法会给出相似的结果。两种估计值都属于最优渐近正态估计值（best asymptotically normal estimators）。但是，基于现实的考虑，往往会选择最大似然估计。例如，空单元格计数往往会影响加权最小二乘法的效果。这种情况下，样本的结果变量函数可能无法清楚界定或者所估计的协方差矩阵可能为奇异阵。

与类似然法相同，加权最小二乘法的推断结果仅取决于设定一个关于结果变量均值的模型，以及指定结果变量的方差函数和协方差结构（这里，是指基于多项分布的情况）。它不需要使用整个分布的似然函数。因此，在进行统计推断时，应当使用沃尔德方法。

在过去，加权最小二乘法的一个优点在于运算简便。现在，随着专门进行最大似然分析的软件的出现，以及加权最小二乘法的扩展方法（如，GEE 等类似然法）改进了其局限性，这一点变得不再重要。因此，与大约 25 年前相比，现在对加权最小二乘法的应用

少了很多。不过,它与一些更完善的方法之间存在着密切联系。一些用于计算最大似然估计的算法也包括对加权最小二乘法的迭代使用。Miller 等(1993)指出,在一定条件下,GEE 拟合过程的第一次迭代所给出的解就是加权最小二乘法估计。这个解直接使用样本值作为初始估计并假定存在饱和的关联结构,即允许结果变量的每对类别以及组内的每对观测值之间都存在不同的相关关系参数。在这个意义上,GEE 是一种迭代形式的加权最小二乘法,而且,在这种情况下,两种方法所给出的估计结果的协方差矩阵也都相同。

15.2 分类数据的贝叶斯推断

在过去十年间,基于贝叶斯理论的统计方法取得了长足进展。在新的运算能力下,估计模型参数的后验分布变得更加容易。然而,对分类数据分析来说,贝叶斯推断的发展与应用不如在其他统计学领域那么普遍。其部分原因在于,在分析多维列联表数据时,多项分布模型包括大量的参数,常常要求进行大量的先验设定。关于贝叶斯理论与方法的探讨超出了本书的范围。在此,我们只介绍一些相对基本的问题,对其应用贝叶斯方法是一种很自然的选择,而且有时候这种方法具有最大似然法所不具备的优点。接着,我们简要综述一些相对复杂的研究进展。

第一种利用贝叶斯方法分析列联表数据的应用是,通过修匀单元格计数来改善对单元格概率的估计(如:Good,1965)。样本比例是饱和模型对应的普通最大似然估计值。在数据稀疏的情况下,这些估计值存在一定的问题。大型稀疏表格往往包含许多抽样性零值(sampling zeros),其概率估计为 0.0,一般不能令人满意。此外,Stein 关于多元正态均值估计的结果表明,将样本比例向某些平均值收缩的贝叶斯估计值具有较小的均方差(mean-squared error)(Efron and Morris,1975)。

在进行贝叶斯估计时,我们不能奢望找到一个在各方面都优于最大似然估计的估计值。例如,假定一个单元格的真实概率 $\pi_i=0$,这时,样本比例 $p_i=0$ 的概率为 1,而且它比任何其他估计值都好。由于存在使样本比例最优化的参数值,没有任何一个估计值在整个参数空间中都优于其他估计值。这里,比较的标准是一个测量估计值与参数之间距离的*损失函数*(*loss function*)的期望值,如平方差(squared error)。按照决策理论的术语,就标准的损失函数而言,样本比例是一个*可容许*(*admissible*)估计值(Johnson,1971)。在这个意义上,多项分布或多元二项分布的样本均值不同于多元正态分布的样本均值,而后者在均值向量的维度大于等于 3 时是不可容许的(由贝叶斯估计值所支配)(Ferguson,1967,p. 170)。Meeden 等(1998)给出了可分解的对数线性模型的相应结果。

另一种估计单元格概率的方法是拟合非饱和模型。但是,通常来说,我们并不期望某个特定模型能很好地描述表格。例如,对于定类变量的 $I\times J$ 交互表格,独立性模型几乎很少能充分拟合数据。当非饱和模型对真实关系的近似很差时,模型估计值也存在很多问题。尽管这些估计值修匀了数据,但是对大样本来说,这是一种过度的修匀。这时,模型估计值不满足一致性,随着 n 的上升,它所收敛的值可能远远不同于真实的单元格概率。

估计单元格概率的贝叶斯方法是对样本比例与模型估计值的一种折中。这时,模型仍然起到了部分的修匀作用,贝叶斯估计值将样本比例向满足模型的一组比例进行收缩。

15.2.1　二项分布参数的贝叶斯估计

我们具体介绍一下对二项分布参数进行贝叶斯推断的基本原理。令 y 表示一个 $\text{bin}(n,\pi)$ 变量，由于 π 的取值落在 0 到 1 之间，所以可以考虑 π 的先验密度函数服从 β(beta)分布（第 13.3.1 节中的式 13.8），其中 $\alpha>0$ 且 $\beta>0$。这个分布满足 $E(\pi)=\alpha/(\alpha+\beta)$。

在贝叶斯推断中，给定数据后参数的后验密度函数与先验密度函数和似然函数之间的乘积成比例。这里，β 先验函数通过 $\pi^{\alpha-1}(1-\pi)^{\beta-1}$ 取决于 π，而且二项分布似然函数的核函数通过 $\pi^{y}(1-\pi)^{n-y}$ 取决于 π。因此，对于 $0\leqslant\pi\leqslant1$，π 的后验密度函数 $h(\pi|y)$ 与

$$h(\pi|y)\propto[\pi^{y}(1-\pi)^{n-y}][\pi^{\alpha-1}(1-\pi)^{\beta-1}]=\pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1}$$

成比例。β 是一个共轭先验分布。后验密度函数也是 β 函数，它的参数分别为 $\alpha^*=y+\alpha$ 和 $\beta^*=n-y+\beta$。

后验分布的均值是关于参数的一个贝叶斯估计值。当平方差损失函数 $(T-\pi)^2$ 能够描述通过估计值 T 来估计 π 的后果时，它是最优的选择（Ferguson，1967，p. 46）。π 的 β 后验分布均值为

$$\begin{aligned}E(\pi\mid y)&=\alpha^*/(\alpha^*+\beta^*)=(y+\alpha)/(n+\alpha+\beta)\\&=w(y/n)+(1-w)[\alpha/(\alpha+\beta)],\end{aligned}$$

其中 $w=n/(n+\alpha+\beta)$。这等于对样本比例 $p=y/n$ 和先验分布均值的加权平均。对于给定的 (α,β)，随着 n 的上升，对样本比例所赋的权重也增大。后验分布的标准差描述了该估计值的精确度，它等于

$$\text{var}(\pi\mid y)=\alpha^*\beta^*/(\alpha^*+\beta^*)^2(\alpha^*+\beta^*+1)$$

的平方根。在大样本的情况下，这个标准差大约等于 $\sqrt{p(1-p)/n}$，即普通最大似然估计值 $\hat{\pi}=p$ 的标准误。

进行贝叶斯估计时，需要选取先验分布中的参数 (α,β)。在完全忽视 π 的情况下，可能会选择均匀先验分布，这相当于 $\alpha=\beta=1$ 时的 β 分布。这时，后验分布具有与二项分布似然函数相同的形状。相应的贝叶斯估计值为

$$E(\pi\mid y)=(y+1)/(n+2)。$$

该估计值将样本比例向 $\frac{1}{2}$ 进行了微小的收缩。

另一种常用的贝叶斯先验分布是 Jeffreys *先验分布*（*Jeffreys prior*），它与所关注参数的 Fisher 信息矩阵行列式（determinant）的平方根成比例。在只有一个参数 θ 的情况下，它等于 $[E(\partial^2\log f(y|\theta)/\partial\theta^2)]^{1/2}$。对于二项分布变量，$\theta=\pi$ 且 $n=1$，上式等于 $[\pi(1-\pi)]^{-1/2}$，即先验分布是当 $\alpha=\beta=0.5$ 时的 β 分布。Brown 等（2001）表明，利用这一先验分布所生成的后验分布中的关于 π 的置信区间令人满意，它近似于经过中位 P 值调整（mid-P adjustment）后的 Clopper-Pearson 区间（第 1.4.4 节和第 1.4.5 节）。在检验备择假设为 $H_a:\pi<\frac{1}{2}$ 的 $H_0:\pi\geqslant\frac{1}{2}$ 时，贝叶斯 P 值等于后验分布中 $\pi\geqslant\frac{1}{2}$ 的概率。Routledge（1994）表明，在使用 Jeffreys 先验分布的情况下，这个后验概率近似等于普通二项分布检验中单边的中位 P 值。

15.2.2　多项分布参数的 Dirichlet 先验/后验分布

上述原理可以从二项分布参数扩展到多项分布参数的情况（Good，1965）。假定单元

格计数$(n_1,\cdots,n_N)$服从$n=\sum n_i$、参数为$\boldsymbol{\pi}=(\pi_1,\cdots,\pi_N)'$的多项分布。多项分布似然函数与

$$\prod_{i=1}^{N}\pi_i^{n_i}$$

成比例。对于$\boldsymbol{\pi}$的可能取值的先验分布,可以考虑β分布的多元扩展——Dirichlet *密度函数*(*Dirichlet density*)

$$g(\boldsymbol{\pi})=\frac{\Gamma(\sum\beta_i)}{\prod_i\Gamma(\beta_i)}\prod_{i=1}^{N}\pi_i^{\beta_i-1},\text{对所有 } i \text{ 存在 } 0\leqslant\pi_i\leqslant 1,\text{且}\sum_i\pi_i=1,$$

其中$\{\beta_i>0\}$。在此函数中,$E(\pi_i)=\beta_i/(\sum_j\beta_j)$。

它的后验分布也是一个 Dirichlet 函数,其参数为$\{n_i+\beta_i\}$。关于π_i的贝叶斯估计值为

$$E(\pi_i\mid n_1,\cdots,n_N)=(n_i+\beta_i)/(n+\sum_j\beta_j)。\tag{15.3}$$

令$K=\sum\beta_j,\gamma_i=E(\pi_i)=\beta_i/K$,$\{\gamma_i\}$是关于单元格概率的先验猜测。贝叶斯估计值等于以下加权平均:

$$[n/(n+K)]p_i+[K/(n+K)]\gamma_i。\tag{15.4}$$

由式 15.3 可得,当先验信息对应着结果分别为β_i的$\sum_j\beta_j$次试验时,$i=1,\cdots,N$,贝叶斯估计值等于样本比例。这种解释对$\{\beta_i\}$的选取具有指导意义。Jeffreys 先验分布设定所有$\beta_i=0.5$。Good 将K称为*扁平常数*(*flattening constant*),因为在所有$\{\beta_i\}$都相等的情况下,式 15.4 将每个样本比例向同一个值$\gamma_i=1/N$进行收缩。对于给定的n,K越大,扁平的程度也就越大。层级模型(hierarchical models)将$\{\beta_i\}$视为未知的,并对其设定二阶先验分布(second-stage prior)(如:Albert and Gupta,1982)。

贝叶斯估计值综合了样本比例和模型估计值的优点。与样本比例相同但不同于模型估计值,即使当模型不成立时,贝叶斯估计值仍然具有一致性。除非模型成立,否则随着样本规模的增加,样本比例的权重上升至 1.0。与模型估计值一样但不同于样本比例,贝叶斯估计值对数据进行了修匀。尽管存在细微的偏差,由此所得的估计值往往具有比样本比例更小的均方差(mean-squared error)。

15.2.3 分类数据贝叶斯方法的发展

现在,我们回顾一下从 Good(1965)用以修匀多项分布比例以来,有关分类数据的贝叶斯方法的发展。Leonard 和 Hsu(1994)对此进行了更为详细的综述,我们的讨论从有关二维列联表的方法开始。

对于 2×2 表格,Altham(1969)通过贝叶斯方法比较了两个独立的二项分布样本的参数。利用π_1和π_2的独立$\beta(\alpha_i,\beta_i)$先验分布,她对备择假设为$\pi_1>\pi_2$的$H_0:\pi_1\leqslant\pi_2$进行了检验。Altham 表明,相应检验的P值为$\pi_1\leqslant\pi_2$的后验分布概率,它可能等于 Fisher 精确检验的单边P值。当使用不当的先验分布$(\alpha_1,\beta_1)=(1,0)$和$(\alpha_2,\beta_2)=(0,1)$时,就会出现上述情况。这反映了对零假设的先验偏好,实际上在规避出现$\pi_1>\pi_2$的可能性。换句话说,Fisher 精确检验对应着一个保守的先验分布。

如果$\alpha_i=\beta_i=\gamma,i=1,2$且$0\leqslant\gamma\leqslant 1$,Altham 表明,贝叶斯$P$值小于 Fisher P值,二者之差不超过观察数据发生的零概率(null probability)。利用$\alpha_i=\beta_i=0.5$的 Jeffreys 先验

分布对 Fisher 精确检验进行连续性修正,这与频数论方法下中位 P 值的作用类似。Howard(1998)指出,在应用这些先验分布时,$\pi_1 \leqslant \pi_2$ 的后验概率近似等于大样本 z 检验的单边 P 值,后者利用联合方差(即皮尔逊统计量的带符号的平方根;参见习题 3.30)对备择假设为 $H_a: \pi_1 > \pi_2$ 的 $H_0: \pi_1 = \pi_2$ 进行检验。Howard 还讨论了有关 2×2 表格的其他先验分布,包括将 π_1 和 π_2 视为相依的情况。

Altham(1971)还对配对数据中的二项分布比例进行了贝叶斯分析。对于在给定时点每个对象的成功概率都相等的简单模型,她再次指出,经典的精确检验的 P 值(第 10.1.4节,利用二项分布)对应于一个偏好 H_0 的先验分布下的贝叶斯 P 值。对于如式 10.8 模型那样允许概率在对象之间变动但时点效应相同的模型,她表明,在两个时点结果不同的配对数保持不变的情况下,随着在两个时点都给出相同结果的配对数的增加,拒绝零假设的贝叶斯证据会变弱。这与条件最大似然估计的结果不同,它不取决于结果相同的配对(第 10.2.3 节)。Ghosh 等(2000)也得出了相似的结论。

到目前为止,本章所介绍的贝叶斯方法都是利用先验分布直接考察单元格概率。Lindley(1964)分析了 $I \times J$ 表格的情况。他考虑了对数概率之间的比较,如对数发生比之比的后验分布。还有一种方法(Laird,1978;Leonard,1975)利用正态的先验分布分析饱和对数线性模型中的参数。正态先验分布并不属于共轭先验分布,但可以通过正态分布来对后验分布进行近似。对关联参数使用独立的正态 $N(0,\sigma^2)$ 分布可以使估计结果向独立性模型收缩(Laird,1978)。在层级方法中,可以对先验分布的参数设定第二阶的先验分布(second-stage priors)(Leonard,1975)。

在过去,使用贝叶斯方法的一个障碍在于,当先验分布不是共轭分布时计算后验分布非常困难。随着通过样本模拟来近似后验分布的现代方法的发展,这不再是一个严重的问题。相应模拟方法包括蒙特卡洛模拟的重要性抽样扩展(Zellner and Rossi,1984)以及马尔科夫链蒙特卡洛法如 Gibbs 抽样(Gelfand and Smith,1990)。Zellner 和 Rossi 在 logistic 回归中使用了贝叶斯方法,Gelfland 和 Smith 则通过 Dirichlet 先验分布讨论了一组多项分布模型。Zeger 和 Karim(1991)利用有关固定效应和随机效应的先验分布,在贝叶斯框架下拟合了广义线性混合模型(GLMMs)。

在广义线性混合模型中对随机效应分布的研究,如 Zeger 和 Karim(1991)等文章,引发了有关广义线性模型的完全贝叶斯方法,即将模型参数视为随机变量。Dey 等(2000)主编的书中收集了一系列利用贝叶斯方法分析广义线性模型的文献。例如,其中 Gelfand 和 Ghosh 对这一主题进行了回顾,Albert 和 Ghosh 综述了项目反应模型,Chib 对不独立的二分数据进行了模型分析,而 Chen 和 Dey 则对相关联的定序数据进行了模型分析。

贝叶斯方法的应用日益广泛。例如,Skene 和 Wakefield(1990)对多中心二分结果变量进行了模型分析,他们使用了一个允许干预方式与结果变量之间的对数发生比之比在各中心之间变动的 logit 模型。除在第 12.3.4 节所介绍的 GLMM 分析以外,它提供了一种贝叶斯分析方法。Daniels 和 Gatsonis(1999)利用多层次广义线性模型分析了具有群组问题的纵向二分数据的空间和时间趋势。他们的研究借鉴了由 Wong 和 Mason(1985)所引入的混合模型的思想。Landrum 和 Normand 的文章(收入在:Dey et al.(2000))给出了使用贝叶斯定序 probit/logit 模型进行分析的例子。Chaloner 和 Larntz(1989)利用贝叶斯方法讨论了如何通过 logistic 回归来决定最优的实验设计。J. Albert 提出,可以通过贝叶斯方法来进行各种分类数据分析。例如,Albert(1997)使用相应方法分析了二维表格的关联,Albert 和 Chib(1993)讨论了二分数据的回归模型,集中介绍了 probit 模型以及有关定序结果变量的扩展。

15.2.4 选取先验分布时的数据依赖

在贝叶斯分析中，有必要细致地考虑如何设定先验分布。使用不适当的先验分布，比如在整个（或正的）实数范围内的均匀先验分布，有可能导致错误的后验分布，而且，贝叶斯拟合软件的输出结果并不会告诉我们这一问题。另外，在使用模拟方法时，拟合过程在什么地方实现了收敛可能并不明显。如果贝叶斯估计结果与频数论的普通最大似然结果相差很大，就需要非常谨慎。

一些统计学家不喜欢贝叶斯方法在选取先验分布时不可避免的主观性。与直接为先验分布选取特定参数不同，越来越流行的做法是使用一种层级方法，即设定这些参数本身服从第二阶先验分布。此外，经验贝叶斯方法利用数据本身来决定先验分布中所使用的参数值（如：Efron and Morris，1975）。该方法通过对先验分布求积分，选取使观察数据的边际概率最大化的先验分布。Laird（1978）在对数线性模型中使用了这一方法，对关联参数的正态先验分布，他通过求解使观察数据中单元格计数的边际分布最大化的值，从而估计先验分布中的 σ^2。与层级方法相比，经验贝叶斯方法的一个缺点在于，它没有考虑由于对先验分布参数使用估计值所导致的变动性。

Fienberg 和 Holland（1973）提出利用具有数据依赖的先验分布来进行列联表分析。对于贝叶斯估计值（式 15.4）中 Dirichlet 均值 $\{\gamma_i\}$ 的特定取值，他们指出，当满足

$$K = \frac{1 - \sum \pi_i^2}{\sum (\gamma_i - \pi_i)^2} \tag{15.5}$$

时，总和均方差（total mean squared error）取最小值。最优的 $K = K(\boldsymbol{\gamma}, \boldsymbol{\pi})$ 取决于 $\boldsymbol{\pi}$，因而他们使用了 K 的估计值 $K(\boldsymbol{\gamma}, \mathbf{p})$，即由样本比例 $\mathbf{p}$ 替代 $\boldsymbol{\pi}$。随着 $\mathbf{p}$ 与先验猜测值 $\boldsymbol{\gamma}$ 不断接近，$K(\boldsymbol{\gamma}, \mathbf{p})$ 的值会上升，在后验估计中对先验猜测所赋的权重也就越大。他们利用一个简单模型的拟合结果，选取了单元格概率的先验分布模式 $\{\gamma_i\}$。在二维表格的情况下，他们使用了独立性模型的拟合结果 $\{\gamma_{ij} = p_{i+}p_{+j}\}$。相应地，贝叶斯估计值将样本比例向独立性模型的拟合值进行收缩。

与其他的统计推断方法相同，在进行贝叶斯分析时应当考虑结果变量类别之间的排序性。例如，在上文刚刚提及的修匀列联表数据的方法中，应当考虑使样本比例向定序模型的拟合值收缩。

15.3 其他估计方法

在本章的最后一节中，我们介绍关于分类数据的一些其他估计方法。假定存在模型 $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$，考虑对 $\boldsymbol{\pi}$ 或 $\boldsymbol{\theta}$ 的估计。令 $\tilde{\boldsymbol{\theta}}$ 表示关于 $\boldsymbol{\theta}$ 的一般性估计值，这里，$\tilde{\boldsymbol{\pi}} = \boldsymbol{\pi}(\tilde{\boldsymbol{\theta}})$ 就是对 $\boldsymbol{\pi}$ 的估计。最大似然估计值 $\hat{\boldsymbol{\theta}}$ 使似然函数取最大值，$\hat{\boldsymbol{\theta}}$ 同时也使对观察比例与拟合比例进行比较的偏离度统计量 G^2 取最小值（第 14.3.4 节）。

15.3.1 最小卡方估计值

与最大似然估计值相似，其他估计值也是使某些测量 $\boldsymbol{\pi}(\boldsymbol{\theta})$ 与 $\mathbf{p}$ 之间距离的其他指标最小化对应的估计值，其中，使皮尔逊统计量

$$X^2[\boldsymbol{\pi}(\boldsymbol{\theta}), \mathbf{p}] = n \sum \frac{[p_i - \pi_i(\boldsymbol{\theta})]^2}{\pi_i(\boldsymbol{\theta})}$$

取最小值的 $\tilde{\boldsymbol{\theta}}$ 被称为*最小卡方*(*minimum chi-squared*)估计。用样本比例替代上式的分母,调整后的统计量为

$$X^2_{\text{mod}}[\boldsymbol{\pi}(\boldsymbol{\theta}),\mathbf{p}] = n\sum \frac{[p_i - \pi_i(\boldsymbol{\theta})]^2}{p_i}, \tag{15.6}$$

对调整后的统计量进行最小化估计在运算上更为简便。相应地,这种估计被称为*最小调整卡方*(*minimum modified chi-squared*)估计,它等于下列方程中关于 $\boldsymbol{\theta}$ 的解:

$$\sum_i \frac{\pi_i(\boldsymbol{\theta})}{p_i}\left(\frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_j}\right) = 0,\quad j = 1,\cdots,q。$$

最小调整卡方估计值是由 Neyman(1949)提出的。他表明,这些估计值与最小卡方估计值都属于最优渐近正态(best asymptotically normal,BAN)估计值。当模型成立时,它们渐近(随着 $n\to\infty$)等价于最大似然估计值。在大样本的情况下,当模型成立时,不同估计方法(最大似然法、加权最小二乘法、最小卡方法等)给出的参数估计结果几乎都相同。这部分是因为估计值满足一致性,随着 n 的上升,它们依概率收敛于 $\boldsymbol{\theta}$。然而,当模型不成立时,即便是在大样本的情况下,不同估计方法所得到的估计值可能也会相差很大。这时,估计值收敛于使模型对现实数据的近似达到最优的值,但是在不同方法中有关最优的定义,如最小化 G^2,X^2 或是其他指标,却是不同的。

对于任意的 n,最小调整卡方估计有时与加权最小二乘法估计完全相同。二者的联系源于模型的另一种设定方式,对 $\boldsymbol{\pi}$ 应用一组*限定方程*(*constraint equations*),

$$\{g_j(\pi_1,\cdots,\pi_N) = 0\}。$$

例如,对于 $I\times J$ 表格,$(I-1)(J-1)$ 个限定方程

$$\log \pi_{ij} - \log \pi_{i,j+1} - \log \pi_{i+1,j} + \log \pi_{i+1,j+1} = 0$$

设定了独立性模型。限定方程的数量等于模型的残差自由度。

Neyman(1949)指出,最小调整卡方估计来自于对 $\boldsymbol{\pi}$ 求

$$\sum_{i=1}^{N}\frac{(p_i - \pi_i)^2}{p_i} + \sum_{j=1}^{N-q}\lambda_j g_j(\pi_1,\cdots,\pi_N)$$

的最小值,其中 $\{\lambda_j\}$ 为拉格朗日算子(Lagrange multipliers)。当限定方程是关于 $\boldsymbol{\pi}$ 的线性方程时,所得到的估计方程也是线性的。Bhapkar(1966)证明,上述情况下的最小调整卡方估计值与加权最小二乘法估计值完全一致。这时,式 15.6 统计量也就等于加权最小二乘法中检验模型拟合的残差统计量(第 15.1.2 节)。

然而,一般来说,限定方程是关于 $\boldsymbol{\pi}$ 的非线性方程,比如独立性模型的情况。这时,加权最小二乘法估计值是最小修正卡方估计值,它基于一种线性化的限定方程,

$$g_j(\mathbf{p}) + (\boldsymbol{\pi} - \mathbf{p})'\partial g_j(\boldsymbol{\pi})/\partial \boldsymbol{\pi} = 0,$$

其中的微分向量对应于在 $\mathbf{p}$ 处的取值。

Berkson(1944,1955,1980)强烈推荐使用最小卡方法。在 logistic 回归中,他的*最小 logit 卡方*(*minimum logit chi-squared*)估计值是对样本 logit 及其线性预测值之间的加权总平方和求最小值。Mantel(1985)则对这些方法提出了批评,指出相应估计值的一致性需要在各组计数都很大时才成立,而最大似然法(或存在许多冗余参数时的条件最大似然法)不管信息多么有限都具有一致性(另见习题 15.14)。

15.3.2　最小判别信息

Kullback(1959)阐述了通过*最小判别信息*(*minimum discrimination information*,*MDI*)来进行估计的方法。两个概率向量 $\boldsymbol{\pi}$ 和 $\boldsymbol{\gamma}$ 的判别信息为

$$I(\boldsymbol{\pi};\boldsymbol{\gamma}) = \sum_{i=1}^{N} \pi_i \log(\pi_i/\gamma_i)。 \tag{15.7}$$

这个对 $\boldsymbol{\pi}$ 和 $\boldsymbol{\gamma}$ 之间距离的直接测量是非负的,且仅当 $\boldsymbol{\pi}=\boldsymbol{\gamma}$ 时它才等于 0。Gokhale 和 Kullback(1978)讨论了最小化 $I(\boldsymbol{\pi};\boldsymbol{\gamma})$ 的最小判别信息估计。对部分模型他们使用的限定条件为 $\boldsymbol{\gamma}=\mathbf{p}$,对其他模型使用的条件则为 $\gamma_1=\gamma_2=\cdots=\gamma_N=1/N$。Good(1963)在*最大熵*(*maximum entropy*)方面进行了有关研究。

在限定条件为 $\{\gamma_i=1/N\}$ 时的部分情况下,最小判别信息估计值与最大似然估计值完全一致(Simon,1973)。当限定条件为 $\boldsymbol{\gamma}=\mathbf{p}$ 时,它与最大似然估计值不同,但二者具有相似的特性,都属于最优渐近正态(best asymptotically normal,BAN)估计值。Gokhale 和 Kullback 建议利用二倍的 $I(\boldsymbol{\pi};\mathbf{p})$ 的最小值来检验模型的拟合优度。与 G^2 相比,这个统计量互换了 $\mathbf{p}$ 和 $\boldsymbol{\pi}$ 的角色,就像在式 15.6 中 X^2_{mod} 互换了二者在 X^2 中的角色一样。两个统计量均属于效能多样化(power divergence)统计量(Cressie and Read,1984;另见习题 3.34),并具有相似的渐近特性。更一般地说,大家可以选取任意一种效能多样化统计量,并界定使其取最小值的估计值。在规则性条件下,它们都是最优渐近正态估计值。

15.3.3 核修匀

核估计(kernel estimation)是一种在不需要做任何参数分布假定的情况下估计概率密度函数的修匀方法。令 $\mathbf{K}$ 表示一个包含非负元素且各列相加等于 1 的矩阵。列联表中单元格概率的核估计具有如下形式:

$$\tilde{\boldsymbol{\pi}} = \mathbf{K}\mathbf{p}。 \tag{15.8}$$

对于具有 N 个类别的定类变量,Aitchison 和 Aitken(1976)使用了

$$K_{ij} = \begin{cases} \lambda, (i=j) \\ (1-\lambda)/(N-1), (i \neq j) \end{cases},$$

其中 $(1/N) \leqslant \lambda \leqslant 1$。由此得到的关于 $\boldsymbol{\pi}$ 的核估计值为

$$(1-\alpha)\mathbf{p} + \alpha(\mathbf{1}/N), \tag{15.9}$$

其中 $\alpha=N(1-\lambda)/(N-1)$。这个估计值将样本比例向 $(1/N,\cdots,1/N)$ 收缩。随着 λ 从 1 下降到 $1/N$,修匀参数 α 从 0 上升到 1。Brown 和 Rundell(1985)证明,当没有 $\pi_i=1$ 时,存在 $\lambda<1$ 使得该核估计值的总和均方差小于样本比例的总和均方差。对多元均值应用其他收缩估计值(shrinkage estimators)的结果表明,在小样本且真实的单元格概率大致相等的情况下,核估计值是对样本比例的明显改进。

Brown 和 Rundell 将核修匀(kernel smoothing)扩展到包含定类和定序变量的多维列联表的情况。在 T 维表格中,令 $\mathbf{L}_k$ 表示一个随机矩阵(即行和列的和等于 1),其元素为

$$l_{k,ij} = \begin{cases} \lambda_k, & i=j \\ d_k(i,j)(1-\lambda_k), & i \neq j \end{cases},$$

$k=1,\cdots,T$。他们令式 15.8 中的 K 等于 Kronecker 乘积(Kronecker product)

$$\mathbf{K} = \mathbf{L}_1 \otimes \cdots \otimes \mathbf{L}_T。$$

当变量 k 是定序变量时,仅仅进行收缩是不够的,还需要借用附近单元格的信息。对于类别之间距离较大的 i 和 j,所选取的 $d_k(i,j)$ 就较小一些。如果变量 k 是定类变量,一种自然的选择为 $d_k(i,j)=1/(I_k-1)$,其中 I_k 是变量 k 所包括的类别数量。对于给定的 $\{\lambda_k\}$,对修匀后的表格进行合并与先将原始表格合并再进行修匀所得到的结果相同。在 $\{\lambda_k=\lambda, k=1,\cdots,T\}$ 的情况下,Brown 和 Rundell 介绍了求解 λ 以得到最小化总和均方差的无偏估计的方法。

Dong 和 Simonoff(1995)以及 Simonoff(1986)描述了有关定序类别的其他方法。大多数情况下,这种核修匀方法得到的概率估计结果具有以下形式

$$\tilde{\pi}_i = (1-\alpha)p_i + \alpha \times 修匀值_i,$$

其中当周边单元格的真实概率相似时,修匀的结果原则上会很好。

15.3.4 惩罚性似然法

Good 和 Gaskins(1971)介绍了用于估计密度函数的*惩罚性似然*(*penalized likelihood*)法。就对数似然函数 $L(\boldsymbol{\pi})$ 而言,惩罚性似然估计值最大化

$$L^*(\boldsymbol{\pi}) = L(\boldsymbol{\pi}) - \alpha(\boldsymbol{\pi}),$$

其中 $\alpha(\cdot)$ 是一个粗糙度惩罚(roughness penalty)函数。也就是说,在一定意义上随着 $\boldsymbol{\pi}$ 的元素越平滑,$\alpha(\boldsymbol{\pi})$ 会下降。惩罚性似然估计值具有贝叶斯意义上的解释。当先验密度函数与 $\exp[-\alpha(\boldsymbol{\pi})]$ 成比例时,后验密度函数就与惩罚性似然函数成比例。因此,后验分布的众数等于惩罚性似然估计值。

Simonoff(1983)利用惩罚性似然法对单元格概率 $\boldsymbol{\pi}$ 进行了估计。与贝叶斯方法和核修匀方法一样,惩罚性似然估计值比样本比例更平滑。对于单个的定序变量,Simonoff(1983)使用了惩罚函数 $\alpha(\boldsymbol{\pi}) = \lambda \sum_{i=1}^{N-1}(\log \pi_i - \log \pi_{i+1})^2$,它使得相邻类别的估计值相似。在二维列联表中,Simonoff 建议使用 $\alpha(\boldsymbol{\pi}) = \lambda \sum_i \sum_j (\log \theta_{ij})^2$,它是关于局部发生比之比的函数。相应的估计结果将样本比例向独立性模型的估计值进行收缩。在选取修匀参数 λ 时,应最小化估计值均方差的近似值。

在对诸如核修匀和惩罚性似然法等修匀方法进行评价时,有必要区分单元格数量 N 给定情况下的大样本渐近特性与 N 随着 n 的增加而增加时的稀疏数据渐近特性(回顾第 6.3.4 节)。在前一种情况下,这些修匀方法以及贝叶斯推断的渐近特性与普通最大似然结果(即样本比例)相似。它们以相同的速度向真实概率收敛。这时,修匀方法对最大似然结果的改进主要体现在小样本的情况,它们具有"借用整体信息"的优点。然而,对于稀疏数据情况下的渐近特性,修匀的优点尤其突出。随着表格维数的增加,单元格数量呈指数增长,从而出现了"维度诅咒(curse of dimensionality)"的问题。随着估计值更缓慢地向真实值收敛,获得精确估计结果变得更加困难。这时,表格中空单元格的比例也会增加,甚至从渐近的角度来说,修匀方法也会优于最大似然法。感兴趣的读者,可以参考以下研究:有关基于 Dirichlet 先验分布的贝叶斯估计值的情况,参见:Fienberg and Holland(1973);有关对多项分布变量进行惩罚性似然估计的情况,参见:Simonoff(1983)。Simonoff 表明,随着 n 和 N 的增加从 $\sup_i |\hat{\pi}_i/\pi_i - 1| \xrightarrow{p} 0$,以及概率本身趋近于 0 的角度来说,惩罚性似然估计值具备一致性。

有关修匀方法的综述,参见:Fahrmeir and Tutz(2001,Chap. 5)、Lloyd(1999,Chap. 5)、Simonoff(1996,Chap. 6;1998)。正如 Simonoff 所指出的,所有修匀方法都试图在修匀不足情况下的低偏差与过度修匀情况下的低变动性之间寻求一个平衡。这些方法需要用户来选取先验分布或者某种形式的修匀参数,从而决定所希望达到的修匀程度。

总之,关于分类数据的修匀方法有很多。除了本节介绍的方法外,还有传统的模型构建方法,甚至有些模型——如广义可加模型(第 4.8 节),也以用于修匀为主要目的。具体到某个应用来说,某种特定修匀方法可能会尤其适用。而贝叶斯方法的一个优点在于,它的整个程式不像其他一些方法那样具有明显的针对性。

注解

第 15.1 节:关于分类数据的加权最小二乘法

15.1 有关加权最小二乘法的应用,包括拟合均值结果变量模型(mean response models)(Grizzle et al.,1969)以及边际分布模型(Koch et al.,1977)。对加权最小二乘法的一般性讨论,参见:Bhapkar and Koch(1968)、Imrey et al.(1981)、Koch et al.(1985)。

第 15.2 节:分类数据的贝叶斯推断

15.2 有关对分类变量进行贝叶斯分析的其他文献包括:Fienberg et al.(1999)、Foster and Smith(1998)、Good(1976)、Knuiman and Speed(1988)、Spiegelhalter and Smith(1982)、Walley(1996)。

第 15.3 节:其他估计方法

15.3 关于最小卡方法的进一步讨论,参见:Bhapkar(1966)、Koch et al.(1985)、Neyman(1949)、Rao(1963)。

15.4 有关最小判别信息的应用,参见:Gokhale and Kullback(1978)、Ireland and Kullback(1968a,b)、Ireland et al.(1969)、Ku et al.(1971)。

15.5 Hall 和 Titterington(1987)探讨了多项分布核估计值的收敛速度,并定义了一个最优速度。普通的核估计值一般在表格的边界上会出现趋于零的偏误。Dong 和 Simonoff(1994)考虑了对大型稀疏表格边界处的核估计的改进问题。核修匀方法还可以应用于离散回归模型分析。在结果变量为二分变量的情况下,Copas(1983)利用一种非参数的核方法展示了 $P(Y=1)$ 对 x 的依赖。

习题

应用部分

15.1 考虑第 7.4.6 节所拟合的均值结果变量模型。展示如何通过加权最小二乘法来完成此分析。分别指出多项分布样本的数量 I,结果变量的类别数量 J,结果变量函数 $\mathbf{F}$,模型矩阵 $\mathbf{X}$,参数向量 $\boldsymbol{\beta}$,以及所估计的协方差矩阵 $\hat{\mathbf{V}}_F$。

15.2 通过加权最小二乘法分析第 11.2.1 节中有关抑郁的跟踪数据。利用统计软件(如 SAS:PROC CATMOD),报告加权最小二乘法的估计结果和标准误,并将该结果与最大似然结果进行比较。

15.3 参考习题 15.2。分析这些数据,并指出:(a)加权最小二乘法与最大似然法的区别;(b)在对多元分类结果变量数据拟合边际模型时,加权最小二乘法与 GEE 法的区别。

15.4 利用第 1.4.3 节的数据,给出关于素食者所占比例的贝叶斯估计。说明你如何选取先验分布,将所得到的结果与最大似然结果加以比较。

15.5 参考表 9.8,假定式 9.12 成立且年龄对呼吸困难和对哮喘的边际效应为线性 logit 关系的模型。

a. 将模型表示为 $\mathbf{C}\log\mathbf{A}\boldsymbol{\pi}=\mathbf{X}\boldsymbol{\beta}$ 的形式,指出其中的 $\mathbf{C}$,$\mathbf{A}$ 和 $\mathbf{X}$。

b. 利用软件拟合该模型,并对估计结果加以解释。

理论与方法

15.6 考虑 $I\times I$ 表格的边际同质性。

a. 令 $\mathbf{F}(\boldsymbol{\pi})=\mathbf{A}\boldsymbol{\pi}$,解释在什么情况下:(i)$\mathbf{F}(\boldsymbol{\pi})=\mathbf{0}$,其中 $\mathbf{A}$ 包括 $I-1$ 行;(ii)$\mathbf{F}(\boldsymbol{\pi})=\mathbf{X}\boldsymbol{\beta}$,其中 $\mathbf{A}$ 包括 $2(I-1)$ 行并且 $\boldsymbol{\beta}$ 包括 $I-1$ 个元素。在第(ii)部

分，给出当 $I=3$ 时的 $\mathbf{A},\boldsymbol{\pi},\mathbf{X}$ 以及 $\boldsymbol{\beta}$。

b. 说明如何通过加权最小二乘法来检验边际同质性（这就是 Bhapkar 检验（式10.16））。

15.7　当加权最小二乘法中的 $\mathbf{F}(\boldsymbol{\pi})=\mathbf{C}[\log(\mathbf{A}\boldsymbol{\pi})]$ 时，证明 $\mathbf{Q}=\mathbf{C}[\mathrm{diag}(\mathbf{A}\boldsymbol{\pi})]^{-1}\mathbf{A}$。

15.8　在加权最小二乘法中，证明 $[\mathbf{F}(\mathbf{p})-\mathbf{X}\boldsymbol{\beta}]'\hat{\mathbf{V}}_F^{-1}[\mathbf{F}(\mathbf{p})-\mathbf{X}\boldsymbol{\beta}]$ 在 $\boldsymbol{\beta}=(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{F}(\mathbf{p})$ 时取最小值。

15.9　结果变量函数 $\mathbf{F}(\mathbf{p})$ 的渐近协方差矩阵为 $\mathbf{V}_F$，推导加权最小二乘法参数估计值 $\mathbf{b}$ 以及预测值 $\hat{\mathbf{F}}=\mathbf{X}\mathbf{b}$ 的渐近协方差矩阵。

15.10　考虑在使用 β 先验分布时关于二项分布参数 π 的贝叶斯估计值。

a. 是否存在一个 β 先验分布，使得所推出的贝叶斯估计值与最大似然估计值恰好相等？

b. 证明对于 β 先验分布的一系列参数值，最大似然估计值是贝叶斯估计值的极限。

c. 找出一个不当的先验密度函数（其积分等于无穷大），使得贝叶斯估计值与最大似然估计值恰好相等（在这个意义上，最大似然估计值是一个*广义贝叶斯估计值*（*generalized Bayes estimator*））。

d. 在利用损失函数 $w(\theta)(T-\theta)^2$ 进行贝叶斯推断时，θ 的贝叶斯估计值等于 $\theta w(\theta)$ 的后验分布期望值除以 $w(\theta)$ 的后验分布期望值（Ferguson，1967，p.47）。当损失函数为 $(T-\pi)^2/[\pi(1-\pi)]$ 时，证明 π 的最大似然估计值等于使用均匀先验分布的贝叶斯估计值。

e. 风险函数等于损失函数的期望，将其视为关于 π 的函数。对于(d)部分的损失函数，证明风险函数等于一个常数（在常数风险的情况下，贝叶斯估计值是*极小化极大值*（*minimax*），其最大风险值不大于所有其他估计值的最大风险值）。

f. 证明 π 的 Jeffreys 先验分布等于当 $\alpha=\beta=0.5$ 时的 β 密度函数。

15.11　对多项分布概率使用 Dirichlet 先验分布，证明 π_i 的后验分布的期望值为式15.3。将这个贝叶斯估计值表达为关于 p_i 和 $E(\pi_i)$ 的加权平均。

15.12　对于式 15.4 贝叶斯估计值，证明总和均方差等于

$$[K/(n+K)]^2[\sum(\pi_i-\gamma_i)^2]+[n/(n+K)]^2(1-\sum\pi_i^2)。$$

证明式 15.5 是使上式取最小值的 K 值。

15.13　参考习题 15.6。在边际同质性的情况下，说明为什么最小修正卡方估计与加权最小二乘法估计完全相同。

15.14　令 y_i 表示关于第 i 组的 $\mathrm{bin}(n,\pi_i)$ 变量，其中 $i=1,\cdots,N$，并且 $\{y_i\}$ 相互独立，考虑模型 $\pi_1=\cdots=\pi_N$，用 π 表示这个共同值。

a. 证明 π 的最大似然估计值为 $p=(\sum_i y_i)/(\sum_i n_i)$。

b. 最小卡方估计值 $\tilde{\pi}$ 是最小化

$$\sum_{i=1}^{N}\frac{[(y_i/n_i)-\pi]^2}{\pi}+\sum_{i=1}^{N}\frac{[(y_i/n_i)-\pi]^2}{1-\pi}$$

的 π 值。上式中的第二项用于比较 $(1-y_i/n_i)$ 和 $(1-\pi)$，即第二个类别所占的比例。如果 $n_1=\cdots=n_N=1$，证明 $\tilde{\pi}$ 使 $Np(1-\pi)/\pi+N(1-p)\pi/(1-\pi)$ 取最小值，并由此证明

$$\tilde{\pi}=p^{1/2}/[p^{1/2}+(1-p)^{1/2}]。$$

注意这个估计值具有趋向$\frac{1}{2}$的偏差。

c. 在所有 $n_i=1$ 的情况下，随着 $N\to\infty$，最大似然估计值具有一致性，而最小卡方估计值却不满足一致性（Mantel，1985），说明理由。

15.15 参考习题 15.14。在 $N=2$ 组且 n_1 和 n_2 为独立观测值的情况下，求关于 $\boldsymbol{\pi}$ 的最小修正卡方估计值，将其与最大似然估计值加以比较。

15.16 证明核估计值（式 15.9）与使用具有 $\{\beta_i=\alpha n/(1-\alpha)N\}$ 的 Dirichlet 先验分布的贝叶斯估计值（式 15.3）相同。利用这一结果，提出一种由数据来决定核估计值中的 α 的方法。

16 分类数据分析的历史回顾*

在本书的最后,我们简要回顾一下分类数据分析方法的发展历史。我们已经看到,分类尺度的变量在社会科学和生物医学研究中随处可见。毫不奇怪,有关分类结果变量的广义线性模型的发展主要是由从事社会科学或生物医学研究的统计学家来完成的。

直到20世纪的最后25年,这些分类数据模型才得到了像连续数据模型在20世纪初期就已经得到的广泛关注。连续变量的回归模型是由 Francis Galton 在1880年代的革命性研究演化发展而来。R. A. Fisher、G. Udny Yule,以及其他统计学家在农业和生物学方面的实验使得回归模型和方差分析(ANOVA)在20世纪中叶就得到了广泛应用。相对而言,尽管早在1900年左右 Karl Pearson 和 Yule 就发表了有关分类变量之间关联的重要文章,但是直到1960年代之前,关于分类结果变量的模型几乎很少受到关注。

分类数据的分析技术在发展之初充满了各种论战。一些重要的统计学家对此做出了开拓性的贡献,但是在不同统计学家之间却往往存在着激烈的意见分歧。

16.1 皮尔逊-尤尔的关联之争

早期分类数据分析方法的发展主要起源于英国,所以我们将历史回顾的第一站放在20世纪初期的伦敦。1900年是一个显而易见的起点,因为在这一年 Karl Pearson 提出了他的卡方统计量(X^2),而 G. Udny Yule 介绍了发生比之比以及相应的关联测量指标。在此之前,许多研究都集中于对较为简单的测量指标的描述。例如,Goodman 和 Kruskal (1959)指出,比利时社会统计学家 Adolphe Quetelet 早在1849年就使用了相对风险这一指标。

在1900年左右,Karl Pearson(1857—1936)已经是统计学界的名人。他当时是伦敦大学学院(University College in London)统计研究室的主任。在此之前的10年中,他的成就包括提出了一族偏斜概率分布(被称为 Pearson *曲线*(*Pearson curves*)),推导出相关系数的乘积矩(product-moment)估计及其标准误,并扩展了 Galton 有关线性回归的研究。事实上,Pearson 是一位真正的全才,其著述所涉猎的领域非常广泛,包括艺术、宗教、哲学、法律、社会主义、女权、物理学、遗传学、优生学以及进化论。皮尔逊提出卡方检验的动机主要有:检验蒙特卡洛轮盘赌的结果到底是不是随机变动的,检查正态分布和 Pearson 曲线对各种数据的拟合程度,以及检验二维列联表的统计独立性。

在20世纪早期,有关分类数据分析的大多文献都致力于如何正确描述关联关系的激烈论战。Pearson 的方法假定二维列联表是由连续的二元分布所构成的(Pearson, 1904,1913)。他主张使用关于潜在连续变量之间关系的近似测量指标,如相关系数。

1904 年,Pearson 引入了术语"*列联*(*contingency*)"以表示"整个分布对独立概率的偏离",并给出了描述偏离程度的测量指标。假定 2 × 2 表格中的计数服从一个潜在的二元正态分布,*四项相关*(*tetrachoric correlation*)是对该分布的相关关系的最大似然估计。它等于二元正态密度函数中的相关系数 ρ,将该密度函数合并成与所观察到的表格具有相同边际比例的 2 × 2 表格后,其生成的单元格概率等于相应的样本比例。*均方列联*(*mean-square contingency*)和*列联系数*(*contingency coefficient*)相当于对 X^2 在(0,1)刻度上进行正态化处理。关于 $I \times J$ 表格的 Pearson 列联系数(习题 3. 33)对 X^2 进行了标准化,使其近似于潜在连续变量之间的相关系数。

George Udny Yule(1871—1951),一位与 Pearson 同时代的英国统计学家,采用了一种不同的方法。在完成了有关多元线性回归以及多元相关系数和偏相关系数的开创性工作后,Yule 在 1900—1912 年间将其关注点转向了列联表中的关联问题。他认为,许多分类变量,比如(注射了疫苗,未注射疫苗)以及(死亡,存活),在本质上就是离散的。Yule 提出了直接应用单元格计数而不需要假定潜在连续变量的测量指标。他推广了发生比之比 θ(据 Goodman(2000)所言,发生比之比的概念可能是由匈牙利统计学家 J. Korosy 首次提出的)并将其转化为 $[-1, +1]$ 的尺度,即 $Q = (\theta - 1)/(\theta + 1)$,现在被称为 Yule 的 Q(习题 2. 36)。针对 Pearson 测量指标需要假定潜在正态性,Yule(Yule, 1912, p. 612)指出:"正态系数最多只能是说给出了一个关于假想变量之间的假设相关系数。就我看来,在科学研究中引入没有必要的、无法验证的假设并不能令人满意。"Yule(1903)还指出,列联表的边际关联与条件关联可能并不一致,后来 E. H. Simpson 对这一问题进行了探讨,即现在被称为*辛普森悖论*(*Simpson's paradox*)。

在 20 世纪的前 25 年中,Karl Pearson 是英国统计学界的领袖,其研究很少受到挑战。Pearson 强势的个性使他难以友好地面对批评性意见,结果他对 Yule 的观点进行了猛烈的抨击。他争论说,Yule 所提出的系数是不适当的。例如,Pearson 声称,该系数的值并不稳定,在以不同方式将 $I \times J$ 表格合并成 2 × 2 表格后,该指标的取值可能会存在很大差异。针对 Yule 的批评,Pearson 和 D. Heron(1913)在《生物计量学》上发表了超过 150 页的激烈回应,Pearson 是该期刊的发起者之一,并且是当时的主编。在一段针对 Yule 广受好评的著作《统计理论导论》(*An Introduction to the Theory of Statistics*)的批评中,他们写道:"如果 Yule 先生的观点被接受,那么现代统计理论的发展将会遭受无法弥补的损害……在(Pearson)所领导下的任何研究,过去从未,以后也绝不会使用(Yule 的 Q)……我们很遗憾地提请大家注意 Yule 先生在关联研究中所走入的歧途,而且我们不得不对他的方法做出批评,这不仅因为 Yule 先生最近对我们的方法所进行的攻击,而且因为一本在许多方面只会将学习统计的学生带入死胡同的教科书对其方法的毫无头脑的称赞。"Pearson 和 Heron 攻击了 Yule 的"半吊子想法"以及"似是而非的推理",并声称如果 Yule"想保住作为一位统计学家的名声",那么他必须放弃他的主张。

现在回顾起来,Pearson 和 Yule 的观点都具有一定的合理性。一方面,对于一些划分尺度来说,比如大多数定类变量,不存在明显的潜在连续分布。另一方面,许多分类变量的应用确实与潜在连续分布具有天然的联系,而且这种联系可以用来作为构建模型和统计推断的指导(如第 7. 2. 3 节)。Goodman(1981a,b)指出,在第 9. 4. 1 和第 9. 6. 1 节所介绍的定序模型是对 Yule 和 Pearson 的观点进行的某种形式的折中,因为当潜在的分布近似为正态分布时,Yule 的发生比之比描述了拟合充分的模型所具有的特征。

在 Pearson-Yule 论战发生的半个世纪之后,Leo Goodman 和 William Kruskal 对列联表关联测量指标的发展进行了综述,并作出了许多他们自己的贡献。他们在 1979 年出版

的书中收录了他们在《美国统计学会期刊》(*Journal of the American Statistical Association*)上发表的四篇非常重要的文章。许多测量指标的提出可以追溯到19世纪。他们在1959年的文章中引述了M. H. Doolittle在1887年的研究,其体现了早期的研究中即便对2×2表格中关联(*association*)的含义都缺乏精确界定:"已知在这样和那样的情况下都成立、这样的情况成立但那样的情况不成立、那样的情况成立但这样的情况不成立,以及这样和那样的情况下都不成立的相应观测值数量,有必要消除各种情况本身所存在的一般量化的相对性,从而确定在这样和那样之间所特定的量化相对性(Having given the number of instances respectively in which things are both thus and so, in which they are thus but not so, in which they are so but not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things)。"Goodman(2000)对此进行了历史回顾,并提出了一种新的测量指标。

16.2 R. A. FISHER 的贡献

Pearson和Yule的争论,与他后期和Ronald A. Fisher(1890—1962)的争论相比,只能算是小巫见大巫。利用一种几何方法,Fisher(1922)引入了*自由度*(*degree of freedom*)来描述卡方分布族的特征。Fisher强调,在对$I\times J$表格进行独立性检验时,X^2具有的自由度为$\mathrm{df}=(I-1)(J-1)$。相反,Pearson(1900,1904)曾论证说,在任何关于X^2的应用中,Fisher后来称之为自由度的指标等于单元格的数量减1,即在二维表的情况下为$IJ-1$。但是,Fisher指出,在利用行和列的估计概率来估计单元格概率的过程中,会导致$(I-1)+(J-1)$个关于拟合值的额外限定条件,因而影响X^2的分布。

不出所料,Pearson(1922)对Fisher认为他的自由度计算公式有误的观点回应以批评。他写道:"我坚持认为(Fisher)这样的观点是完全错误的,作者在《皇家统计学会杂志》(*Journal of Royal Statistical Society*)上散布这种观点对统计学没有任何帮助……我相信我的批评本身会原谅我将其比作要挑战风车的堂吉诃德;他一定是或者毁了自己,或者毁了整个概率误差的理论,因为它们总是基于那些总体未知的样本值。"Pearson宣称,利用行和列的样本比例去估计未知概率对大样本分布几乎没有影响,尽管他自己(Pearson, 1917)也意识到,当单元格计数存在线性限定时必须要调整自由度。Fisher无法再在皇家统计学会(Royal Statistical Society)发表他对Pearson的反驳,他最终只得退出了该学会。

统计学界很快就认识到Fisher是正确的,但是他在与Pearson的各种论战中仍然尝尽艰辛。后来,在为自己的一卷著作集所写的前言中,Fisher回忆道,他1922年的文章"在发表之前必须要经过那些评论家,首先他们不相信Pearson的研究还需要修正,即便存在这种可能性,他们也确信修正的人应当是他们自己"。关于Pearson,Fisher写道:"如果对别人的自由观点回应以暴躁和偏狭是一种衰老的征兆,那么他在很年轻的时候就已经出现了这种征兆。"Fisher(1926)利用Karl Pearson的儿子——E. S. Pearson在独立性假定下所随机生成的11 688个2×2表格,对Pearson家族给予了更沉重的打击。Fisher表明,这些表格的X^2样本均值等于1. 000 01,与Pearson的$IJ-1=3$相比,他关于$\mathrm{df}=(I-1)(J-1)=1$的$E(X^2)$公式所预测的值1. 0要接近得多。Fisher的女儿,Joan Fisher Box(1978),记录了Fisher与Pearson之间包括此在内的各种论战。Hald(1998,pp. 652-663页)、Plackett(1983),以及Stigler(1999,Chap. 19)综述了二人有关卡

方的争论。

Fisher 在当今统计学家中的卓著声誉主要来自于他的理论工作(引入了诸如充分性、信息,以及最大似然估计值的最优特性等概念)以及他在实验设计和方差分析方面的方法论贡献。尽管在分类数据分析方面 Fisher 的工作不是那么广为人知,他仍然做出了相当重要的贡献。另外,他还在自己的研究中很好地应用了这些方法。例如,Fisher 同时也是一位知名的遗传学家。在他的一篇文章中,他使用 Pearson 的拟合优度检验来检查 Mendel 有关自然遗传的理论,他发现,该拟合结果实在太好了,而不可能是真实的(第1.5.3节)。

Fisher 认识到大样本方法在实验室研究中的局限性,他站在了当时提倡专门的小样本方法的最前沿。在他的经典著作《统计研究方法》(*Statistical Methods for Research Workers*)第一版的前言中,他就大样本方法写道:"传统的统计分析体系与现实研究的需要完全不相称。传统方法不仅是用大炮来打麻雀,而且它还没打中!对于简单的实验数据来说,基于无限大样本理论的精巧方法不够精确。只有根据事物本身的特点系统地解决小样本问题,才有可能对现实数据进行精确检验。"Fisher 是最早提倡使用 W. S. Gosset(笔名"Student")有关 t 分布的成果的统计学家之一。《统计研究方法》的第五版(1934)引入了关于 2 × 2 列联表的 Fisher 精确检验。在他 1935 年的著作《实验设计》(*The Design of Experiments*)中,Fisher 描述了他在洛桑实验站工作时,由下午茶歇的经历所激发的品茶实验(第 3.5.2 节)。

到 20 世纪 30 年代中期,终于出现了一些有关分类结果变量的模型构建。紧随 J. H. Gaddum 在 1933 年有关计量结果变量方法(quantal response methods)的报告,Chester Bliss(1934,1935)在毒理学研究中推广了有关二分结果变量的 probit 模型。Bliss 引入了术语 *probit*,但他使用的是均值为 5(而不是 0,目的是为了避免负值)、标准差为 1 的正态累积分布函数的反函数。在 Bliss(1935)的附录中,Fisher(1935b)列出了一种求解模型参数的最大似然估计的算法。该算法与 Newton-Raphson 算法相似,它利用的是期望信息,现在一般被称为 Fisher *计分法*(*Fisher scoring*,第 4.6.2 节)。Stigler(1986,p. 246)和 Finney(1971)将利用正态累积分布函数的反函数对比例进行转换的首次应用归功于德国物理学家 Gustav Fechner 在 1860 年的著作 Elemente der Psychophysik。有关 probit 模型的发展历史,参见:Finney(1971)、McCulloch(2000)。

关于列联表中的同质性关联(homogeneous association,不存在交互效应)的定义最早出现在由英国统计学家 Maurice Bartlett(1935)所写的一篇有关 2 × 2 × 2 表格的文章。Bartlett 给出了如何在满足给定第三个变量的不同取值后另两个变量的发生比之比相等的情况下,求解关于单元格概率的最大似然估计。他将这一贡献归功于 Fisher。

1940 年,Fisher 提出了列联表的典型相关(canonical correlation)方法。他给出了如何对列联表的行和列进行赋值以使得相关系数最大化。他的研究与后来*对应分析*(*correspondence analysis*)方法的发展具有密切联系,后者主要是在法国完成的(如:Benzecri,1973)。

在关于现代统计科学的应用方面,R. A. Fisher 是最重要的统计学家。他女儿为他所写的自传(Box,1978)精彩地记录了他在统计学和遗传学方面所作的重要贡献。Fienberg(1980)则总结了 Fisher 对分类数据分析方法的贡献。

16.3　Logistic 回归

在回归和方差分析中，Bartlett(1937)利用 $\log[y/(1-y)]$ 对观测值 y 为连续比例的情况进行转换(习题6.33)。在1938年发表的一本关于统计表的书中，R. A. Fisher 和 Frank Yates 指出，在分析二分数据时，这个转换可以应用于二项分布参数。1944年，物理学家兼统计学家 Joseph Berkson 引入了术语 Logit，用来表示上述转换。Berkson 表明，Logit 模型的拟合结果与 probit 模型相似，此后，他在推广 logistic 回归方面做了大量的后续工作。1951年，另一位具有很强医学背景的统计学家，Jerome Cornfield，利用发生比之比来近似个案-控制研究中的相对风险。Dyke 和 Patterson(1952)则首次在 Logit 模型中加入了分类预测变量。

David R. Cox 爵士通过他1958年的文章以及1970年的著作《二分数据分析》(*The Analysis of Binary Data*)，让许多统计学家了解了 logistic 回归。大约在同一时期，一篇由丹麦统计学家兼数学家 Georg Rasch 所发表的文章引发了有关项目反应模型的大量研究。这其中最重要的就是包括对象参数和项目参数的 Logit 模型，现在被称为 Rasch *模型*(*Rasch Model*，第12.1.4节)。这些研究在北欧(尤其是丹麦、荷兰和德国)的心理测量学领域影响非常深远，并引发了美国在教育测试领域的诸多进展。

在1970年以前，有关将 logistic 回归扩展到多类别结果变量的情况非常少见(如：Mantel, 1966)，但在大约1970年之后，这方面取得了许多实质性的进展。关于定类结果变量的早期研究主要集中于经济计量学领域。参见：Bock(1970)，McFadden(1974)，Nerlove 和 Press(1973)，Theil(1969，1970)。由于在1970和1980年代关于离散选择模型(第7.6节)的研究，Daniel McFadden 在2000年获得了诺贝尔经济学奖。有关针对定序结果变量的累积 Logit 模型，参见：Bock and Jones(1968)、Simon(1974)、Snell(1964)、Walker and Duncan(1967)、Williams and Grizzle(1972)。基于潜在正态结果变量的累积 probit 模型的历史则要更长一些，参见如：Aitchison and Silvey(1957)，Bock and Jones(1968，Chap. 8)。随着 McCullagh(1980)给出了有关求解累积连结模型最大似然拟合的 Fisher 计分法，累积 Logit/probit 模型得到了越来越广泛的关注。

接下来，关于 logistic 回归的重要进展是其在个案-控制研究中的应用(如：Breslow, 1996；Mantel, 1973；Prentice, 1976a；Prentice and Pyke, 1979；另见第5.1.4节)以及在具有大量冗余参数的研究中用于拟合模型的条件最大似然法(Breslow et al., 1978；Breslow, 1976，1982；Breslow and Day, 1980；Breslow and Powers, 1978；Cox, 1970；Farewell, 1979；Prentice, 1976a；Prentice and Breslow, 1978；Zelen, 1971；另见第6.7节和第10.2节)。此后，条件法也被应用于小样本精确推断的问题(Hirji et al., 1987；Mehta and Patel, 1995；另见第6.7节)。

Nathan Mantel，一个在上文中出现过的名字，对分类数据分析的发展做出了方方面面的贡献。尽管他以1959年提出的 Mantel-Haenszel 检验以及相应的发生比之比估计值所为人熟知，他的研究还涉猎了趋势检验(1963)、多项分布 Logit 和对数线性模型(1966)、个案-控制研究中的 logistic 回归(1973)、具有固定边际的列联表(Gail and Mantel, 1977)、方形列联表分析(Mantel and Byar, 1978)，以及最小卡方估计和沃尔德检验的问题(1985，1987a)。

最近，对 logistic 回归的主要关注点是在群组数据中对存在相关关系的结果变量拟合 logistic 模型，其中，一个分支是关于跟踪数据的边际模型(Diggle et al., 2002；Liang and

Zeger,1986;Liang et al.,1992)。这些文献主要讨论类似然方法,如广义估计方程(generalized estimating equations,GEE)等。另一个分支则是广义线性混合模型(如:Breslow and Clayton,1993)。

在过去的半个世纪中,也许最重要的贡献就是由英国统计学家 John Nelder 和 R. W. M. Wedderburn 在 1972 年所引入的*广义线性模型*(*generalized linear models*)的概念。它将关于二项分布数据的 logistic 和 probit 模型、关于泊松分布数据的对数线性模型,以及历史悠久的关于正态结果变量的回归和方差分析模型都联系在一起。有趣地是,他们用来拟合广义线性模型的算法是 Fisher 计分法,由 R. A. Fisher 在 1935 年为 probit 模型的最大似然拟合而提出。McCulloch(2000)回顾了从 probit 模型到广义线性模型的发展历程,以及有关诸如类似然法等的进一步扩展。

16.4 多维列联表与对数线性模型

在第二次世界大战之后的 25 年间,有关列联表的模型分析取得了长足发展。H. Cramér(1946)推导了参数估计值大样本分布的一般表达式。C. R. Rao(1957,1963)作出了相应的贡献。

在 1949 年,伯克利的统计学家 Jerzy Neyman 提出了*最优渐近正态*(*best asymptotically normal*,BAN)估计值的概念,在此之前,他已经与 E. S. Pearson 一起在假设检验和区间估计方面作出了根本性贡献。这些估计值具有与最大似然估计值相同的最优大样本特性。最优渐近正态估计值包括在比较观察比例与模型预测比例时使类似于卡方的统计量取最小值的估计值(第 15. 3. 1 节)。这类估计值本身包括某些加权*最小二乘法*(*weighted least squares*,WLS)估计值。在现代运算能力提高之前,与最大似然估计相比,加权最小二乘法在计算上的简便是很重要的现实因素。Neyman(1949)在唯一提及 Fisher 的地方暗示 Fisher 没有认识到除了最大似然估计外,其他估计值也可能是最优渐近正态估计值。他写道:“结果……与 R. A. Fisher 的论断相矛盾,其观点‘最大似然方程可能确实由线性频数以及所有 θ 取值有效性的条件推导而来’本身有些含糊。”当然,Fisher 对此进行了回应。比如,在写一个关于 2 ×2 表格的无条件检验的项目方案(1956)时,他写道:“Neyman 和 Pearson 的‘假设检验理论’的原理很容易误导那些追随者做太多的无用功。”

在 1950 年代初期,William Cochran 发表了有关分类数据分析的各种重要问题的研究成果。Cochran 出生于苏格兰,但他的主要生涯是在美国的大学中度过的,包括依阿华州立大学、北卡罗来纳州立大学、约翰霍普金斯大学,以及哈佛大学。他(1940)利用方差恒定转换对泊松分布和二项分布结果变量进行了模型分析。他(1943)认识到了过度离散的问题,并就有关解决办法进行了探讨。在比较多个匹配样本中的比例时,他(1950)提出了一般化的 McNemar 检验(Cochran 的 Q)。Cochran 在 1954 年发表的经典文章中,既介绍了新的方法,又对应用统计学家给出了相应的建议。文章总结了能使 X^2 统计量的卡方近似达到较好结果所需样本规模的一般原则,它还强调了在推断时使用更具体的备择假设(如单一自由度)以及将卡方统计量分割成不同组成部分的重要性。其中的一个例子是 Cochran 所提出的关于几个 2 × 2 表格的条件独立性检验,这与 Mantel 和 Haenszel(1959)的检验有着密切的联系(第 6. 3. 2 节)。另一个例子是在行变量具有定量含义的 $I\times 2$ 表格中检验比例的线性趋势(第 5. 3. 5 节)。另见:Cochran(1955)。Fienberg(1984)就 Cochran 对分类数据分析的贡献做了回顾。

Bartlett 有关 2 × 2 × 2 列联表交互结构的研究,在发表后的 20 年内几乎毫无影响。

事实上,Lancaster(1951)在介绍如何分割 2 × 2 × 2 表格的 X^2 时指出,“毫无疑问,几乎不会用到超过三维的交叉划分”。但是,到 1950 年代中期和 1960 年代早期,Bartlett 的研究被以各种形式扩展到多维表格的情况,参见如:Darroch(1962),Good(1963),Goodman(1964b),Plackett(1962),Roy and Kastenbaum(1956),Roy and Mitra(1956)。这些文章再加上 Martin W. Birch(1963,1964a,b,1965)的重要文章,为 1965—1975 年间有关对数线性模型的研究工作提供了源泉。Birch 的研究成果是他在格拉斯哥大学从未提交的博士论文的一部分。他给出了在各种不同条件下,如何求得三维表格中单元格概率的最大似然估计,并证明了这些最大似然估计在泊松分布样本和多项分布样本中的等价性。此外,他(另见:Watson (1959))还扩展了 Cramér 和 Rao 关于列联表模型大样本分布的理论结果。Mantel(1966)讨论了早期的研究成果,并明确给出了对数线性模型的公式。一篇由法国统计学家 Henri Caussinus(1966)写的综述文章(部分基于其博士论文),对当时分类数据分析的发展状况进行了很好的总结,其中,Caussinus 还介绍了关于方形表格的准对称性模型。

接下来数十年间,关于对数线性模型以及相应 Logit 模型的研究进展主要发生在美国三所大学中:芝加哥大学、哈佛大学,以及北卡罗来纳大学。在芝加哥大学,Leo Goodman 发表了一系列具有开创性的文章,主题涉及卡方的分割、方形表模型(如准独立性模型)、Logit 和对数线性模型的逐步构建方法、推导对数线性参数最大似然估计的渐近方差、潜类模型、关联模型、相关模型,以及对应分析。对其早期研究成果的综述,参见:Goodman(1968, R. A. Fisher 纪念讲座,1970)。有关其后期的工作,参见:Goodman(1985,1996,2000)。Goodman 还在社会科学期刊上发表了一系列文章,对对数线性模型和 Logit 方法的推广和应用产生了广泛影响(如:Goodman, 1969b)。

在过去的 50 年间,就分类数据分析方法的进展来说,Goodman 是最多产的一位。这个研究领域也从他的持续和惊人的研究成果中受益颇多。另外,Goodman 在芝加哥大学的一些学生也作出了重要贡献。1970 年,Shelby Haberman 完成了他的博士论文(其 1974a 著作的基础),对对数线性模型作出了重要的理论贡献。他讨论的主题包括残差分析、最大似然估计的存在性、定序变量的对数线性模型,以及有关参数数量随样本规模增加而增加的模型(如 Rasch 模型)的理论结果。Clifford Clogg 追随了 Goodman 的步伐,在关联模型、人口学、关于比率的模型、普查,以及其他许多方面对社会科学和统计学产生了重要影响。

与 Goodman 的研究同期,有关对数线性/Logit 模型的最大似然方法的研究在哈佛大学展开,主要贡献者为 Frederick Mosteller 的学生(如 Stephen Fienberg)以及 William Cochran。这项研究是由在全国氟烷研究(National Halothane Study)中分析大型多元数据时所遇到的问题(Bishop and Mosteller, 1969;另见在《统计科学》(*Statist. Sci.*)14,1999 第 345 页对 Lincoln Moses 的采访)而激发的。该研究考察氟烷与其他麻醉剂相比,是否更可能导致由于肝脏损伤而引发的死亡。在美国统计学会的主席致词中,Mosteller(1968)描述了早期通过对数线性模型来修匀多维离散数据的应用。Fienberg 及其学生进一步推动了这项研究。由他、Yvonne Bishop 和 Paul Holland 在 1975 年出版的里程碑式的著作,《离散多元分析》(*Discrete Multivariate Analysis*),起到了将对数线性模型介绍给一般统计学界的主要作用,直到现在,该书仍然是一本很好的文献。

在北卡罗来纳大学的研究由 Gary Koch 以及他的几个学生和合作者来完成,他们的研究对生物医学产生了深远影响。他们发展了关于分类数据模型的加权最小二乘法(第 15.1 节)。由 Koch、J. Grizzle 和 F. Starmer 在 1969 年发表的文章推广了这一方法。在


Karl pearson


G.Udny Yule


Ronald A.Fisher


Leo Goodman

图 16.1 分类数据分析发展历史中的四位领袖人物

后来的文章中,Koch 及其同事对该方法进行了大量扩展,用以解决各种问题,其中包括一些对最大似然法来说非常棘手的问题,如对重复测量的分类数据的分析(Koch et al.,1977)。Vasant Bhapkar 在 1966 年表明,在多数情况下,加权最小二乘法估计值与 Neyman 的最小修正卡方估计值完全一致。

早期有关对数线性模型的文献将所有的变量都视为定类变量。Haberman(1974b)和 Simon(1974)介绍了如何在对数线性模型中利用定序尺度的信息。Leo Goodman(1979a,1981a,b,1983,1985,1986)在多篇文章中对该项工作进行了扩展。这些扩展包括了在对数线性模型中通过参数来替代定序赋值的关联模型。Goodman(1985,1986,1996)还探讨了有关的相关模型,并从模型的角度介绍了与之密切相连的对应分析。

具有条件独立性结构的对数线性模型为列联表提供了一种图示模型(graphical models)。它与第 9.1 节所介绍的关联图有关。Darroch 等(1980)为这方面的研究奠定了基础。

16.5 最新的发展(及展望?)

在过去十年间,关于分类数据分析方面最活跃的新领域是有关群组数据的模型分析,比如在跟踪调查以及其他形式的重复测量中所出现的情况。针对同一群组内的结果

变量存在相关关系的问题,现在可以通过多种模型分析的手段来处理。

如同在第 11 章和第 12 章所讨论的那样,对于这类模型很难进行最大似然估计。例如,就复杂的广义线性混合模型来说,很难对回归参数和方差构成同时进行很好的估计。通过积分去掉随机效应部分获得似然方程,需要利用到如数值积分等近似方法。毫不奇怪,在这方面,各种蒙特卡洛方法的应用正与日俱增。其中一种很有前途的方法是蒙特卡洛 EM 算法,该方法在 E 步骤中使用了蒙特卡洛近似(Booth and Hobert, 1999)。这样,在每次迭代中,可以估算蒙特卡洛误差,并且随着迭代过程增加到足够多,我们可以获得与最大似然估计相同的精确结果。

关于群组相关数据的模型分析很可能是未来几年研究的活跃领域。可以肯定,在广义线性混合模型方面将会产生一些重要的研究成果以及进一步的扩展,其中一个重要扩展是广义*可加混合模型*(generalized *additive* mixed models)。到目前为止,有关分类结果变量的时间序列模型还未受到足够重视。对于所有这些结果变量存在相关关系的模型,有关的模型诊断技术极其重要,我们在这方面还有待突破。在跟踪调查研究中,缺失数据是一个常见的问题。这一领域的研究目前相当活跃。

最近的另一个重要进展是,对精确小样本方法的有效算法的发展。通过这些方法,我们可以确保检验的误差不会超出事先设定的水平、置信区间所涵盖的概率至少等于其名义水平。这里,"精确性(exactness)"仅仅是指在推断过程中所使用的概率分布不依赖于任何未知参数。有关的精确方法并不唯一,其中某些方法可能会由于离散性问题而使得结果非常保守。目前,大多数文献关注的是条件法,即以参数的充分统计量为条件而消除冗余参数。因此,条件法的基本思路起源于 Fisher 的精确检验。条件法非常灵活,适用于所有使用典型连结的指数族线性模型,如关于泊松结果变量的对数线性模型以及关于二项分布结果变量的 Logit 模型。有关精确条件方法在运算方面的进展主要体现在由 Cyrus Mehta、Nitin Patel 及其在哈佛大学的同事所发表的一系列文章(如:Mehta and Patel, 1983)中,他们使用了网络算法(network algorithm)。参见 Agresti(1992)、Mehta(1994)、Mehta 和 Patel(1995)所做的有关综述,以及 *StatXact* 和 *LogXact* 的软件手册(Cytel Software,Cambridge,MA,由 Mehta 和 Patel 创建)。

尽管在"精确"方法方面已经取得了长足进展,某些分析仍然无法进行,而且很可能在短时间内也无法解决,其原因在于随着表格规模或样本规模的增加,所需要的计算时间呈指数增加。因此,各种对精确方法进行准确近似的方法与日俱增,其中包括简单蒙特卡洛法(如:Agresti et al., 1979),进行重要性抽样的蒙特卡洛法(如:Booth and Butler, 1999;Mehta et al., 1988),马尔科夫链蒙特卡洛法(MCMC)(Forster et al., 1996),鞍点近似法(saddlepoint approximations)(Pierce and Peters, 1992,Strawderman and Wells, 1998),以及受离散性影响较小的*近似条件方法*(*approximate* conditioning approach)(Pierce and Peters, 1999)。

最后,关于分类数据分析的贝叶斯方法正在成为一个日益活跃的领域。参数之间的相乘特点使贝叶斯模型分析变得更加复杂。有关对概率进行贝叶斯估计的早期应用,参见:Good(1965)、Lindley(1964)。Good(Good,1965)的文章显然源于第二次世界大战期间他与 Alan Turing 在英国布莱切利公园(Bletchley Park)所进行的破译纳粹密码的工作。有关贝叶斯方法在分类数据分析方面的发展,我们已经在第 15.2.3 节介绍过。

对未来进行预测总是充满风险。然而,未来的研究很可能集中于一些运算复杂的方法,如广义线性混合模型。另一个基本上游离于传统模型分析之外的热门主题是,对拥有大量变量的巨型数据运算方法的发展。这些方法——常被称为*数据挖掘*(*data*

mining)——用于处理复杂的数据结构,在牺牲数据结构的简单性和可解释性的同时追求预测能力。数据挖掘的重要应用领域包括遗传学(比如在极高维的列联表中分析离散的DNA序列)以及商业研究(如信用评分和预测消费者行为的树状结构方法)。

本章所做历史回顾的资料来源包括:Stigler(1986)、由 E. S. Pearson 和 M. G. Kendall 主编的《概率论和统计学发展史》(*Studies in the History of Probability and Statistics*)(London:Griffin,1970)、以及数年来与多位统计学家的私人交流,其中包括 Erling Anderson、R. L. Anderson、Henri Caussinus、William Cochran、Sir David Cox、John Darroch、Leo Goodman、Gary Koch、Frederick Mosteller、John Nelder、C. R. Rao、Stephen Stigler、Geoffrey Watson,以及 Marvin Zelen。对于那些已经读到这里的读者,感谢你们的坚持!要想对分类数据分析的发展历史做更全面的了解,可以参阅以下 25 个按照时间顺序列出的文献。这些文献展现了分析方法的发展演变过程。或者,读者也可以参阅有关这一主题的一些早期著作,比如 A. E. Maxwell 的《定性数据分析》(*Analysing Qualitative Data*)(New York:Methuen,1961)、R. L. Plackett 的《分类数据分析》(*The Analysis of Categorical Data*)(London:Griffin,1974),以及 Bishop、Fienberg 和 Holland 的《离散多元分析》(Discrete Multivariate Analysis)(Cambridge,MA:MIT Press 1975)。

Pearson(1900)
Yule(1912)
Fisher(1922)
Bartlett(1935)
Berkson(1944)
Neyman(1949)
Cochran(1954)
Goodman and Kruskal(1954)
Roy and Mitra(1956)
Cox(1958a)
Mantel and Haenszel(1959)
Birch(1963)
Birch(1964b)
Caussinus(1966)
Goodman(1968)
Mosteller(1968)
Grizzle et al.(1969)
Goodman(1970)
Haberman(1974a)
Nelder and Wedderburn(1972)
McFadden(1974)
Goodman(1979a)
McCullagh(1980)
Liang and Zeger(1986)
Breslow and Clayton(1993)

参考文献

Adelbasit, K. M., and R. L. Plackett. 1983. Experimental design for binary data. *J. Amer. Statist. Assoc.* 78: 90-98.

Agresti, A. 1984. *Analysis of Ordinal Categorical Data.* New York: Wiley.

Agresti, A. 1992. A survey of exact inference for contingency tables. *Statist. Sci.* 7: 131-153.

Agresti, A. 1993. Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonal parameters. *Scand. J. Statist.* 20: 63-71.

Agresti, A. 1997. A model for repeated measurements of a multivariate binary response. *J. Amer. Statist. Assoc.* 92:315-321.

Agresti, A. 1999. On Logit confidence intervals for the odds ratio with small samples. *Biometrics* 55: 597-602.

Agresti, A. 2001. Exact inference for categorical data: Recent advances and continuing controversies. *Statist. Medic.* 20: 2709-2722.

Agresti, A., and B. Caffo. 2000. Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures. *Amer. Statist.* 54: 280-288.

Agresti, A., and B. A. Coull. 1998. Approximate is better than exact for interval estimation of binomial parameters. *Amer. Statist.* 52: 119-126.

Agresti, A., and J. Hartzel. 2000. Strategies for comparing treatments on a binary response with multi-centre data. *Statist. Medic.* 19(8): 1115-1139.

Agresti, A., and J. Lang. 1993a. A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika* 80: 527-534.

Agresti, A., and J. Lang. 1993b. Quasi-symmetric latent class models, with application to rater agreement. *Biometrics* 49:131-139.

Agresti, A., and I. Liu. 1999. Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics* 55: 936-943.

Agresti, A., and Y. Min. 2001. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* 57:963-971.

Agresti, A., and R. Natarajan. 2001. Modeling clustered ordered categorical data: A survey. *Internal. Statist. Rev.* 69:345-371.

Agresti, A., D. Wackerly, and J. Boyett. 1979. Exact conditional tests for cross-classifications: Approximation of attained significance levels. *Psychometrika* 44: 75-84.

Agresti, A., C. Chuang, and A. Kezouh. 1987. Order-restricted score parameters in association models for contingency tables. *J. Amer. Statist. Assoc.* 82: 619-623.

Agresti, A., C. R. Mehta, and N. R. Patel. 1990. Exact inference for contingency tables with ordered categories. *J. Amer. Statist. Assoc.* 85: 453-458.

Agresti, A., J. Booth, J. Hobert, and B. Caffo. 2000. Random-effects modeling of categorical response data. *Sociol. Methodol.* 30: 27-81.

Aitchison, J., and C. G. G. Aitken. 1976. Multivariate binary discrimination by the kernel method. *Biometrika* 63: 413-420.

Aitchison, J., and C. H. Cho. 1989. The multivariate Poisson-log normal distribution. *Biometrika* 76: 643-653.

Aitchison, J., and S. M. Shen. 1980. Logistic-normal distributions: Some properties and uses. *Biometrika* 67: 261-272.

Aitchison, J., and S. D. Silvey. 1957. The generalization of probit analysis to the case of multiple responses. *Biometrika* 44: 131-140.

Aitchison, J., and S. D. Silvey. 1958. Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* 29: 813-828.

Aitkin, M. 1979. A simultaneous test procedure for contingency table models. *Appl. Statist.* 28: 233-242.

Aitkin, M. 1980. A note on the selection of log-linear models. *Biometrics* 36: 173-178.

Aitkin, M. 1999. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 55: 117-128.

Aitkin, M., and D. Clayton. 1980. The fitting of exponential, Weibull, and extreme value distributions to complex censored survival data using GLIM. *Appl. Statist.* 29: 156-163.

Aitkin, M., and M. Stasinopoulos. 1989. Likelihood analysis of a binomial sample size problem. Pp. 399-411 in *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, ed. L. J. Gleser, M. D. Perlman, S. J. Press, and A. R. Sampson. New York: Springer-Verlag.

Aitkin, M., D. Anderson, and J. Hinde. 1981. Statistical modelling of data on teaching styles. *J. Roy. Statist. Soc. Ser. A* 144: 419-461.

Aitkin, M., D. Anderson, B. Francis, and J. Hinde. 1989. *Statistical Modeling in GLIM.* Oxford: Clarendon Press.

Albert, J. H. 1997. Bayesian testing and estimation of association in a two-way contingency table. *J. Amer.*

Statist. Assoc. 92: 685-693.

Albert, A., and J. A. Anderson. 1984. On the existence of maximum likelihood estimates in logistic models. *Biometrika* 71: 1-10.

Albert, J. H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* 88: 669-679.

Albert, J. H., and A. K. Gupta. 1982. Mixtures of Dirichlet distributions and estimation in contingency tables. *Ann. Statist.* 10: 1261-1268.

Allison, P. D. 1999. *Logistic Regression Using the SAS System*. Cary, NC: SAS Institute.

Altham, P. M. E. 1969. Exact Bayesian analysis of a 2 × 2 contingency table and Fisher's "exact" significance test. *J. Roy. Statist. Soc. Ser B* 31: 261-269.

Altham, P. M. E. 1970. The measurement of association of rows and columns for an $r \times s$ contingency table. *J. Roy. Statist. Soc. Ser B* 32: 63-73.

Altham, P. M. E. 1971. The analysis of matched proportions. *Biometrika* 58: 561-576.

Altham, P. M. E. 1975. Quasi-independent triangular contingency tables. *Biometrics* 31: 233-238.

Altham, P. M. E. 1978. Two generalizations of the binomial distribution. *Appl, Statist.* 27: 162-167.

Altham, P. M. E. 1984. Improving the precision of estimation by fitting a model. *J. Roy. Statist. Soc. Ser B* 46: 118-119.

Amemiya, T. 1981. Qualitative response models: A survey. *J. Econom. Literature* 19: 1483-1536.

Andersen, E. B. 1970. Asymptotic properties of conditional maximum-likelihood estimators. *J. Roy. Statist. Soc. Ser B* 32: 283-301.

Andersen, E. B. 1980. *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.

Andersen, E. B. 1995. Polytomous Rasch models and their estimation. Pp. 272-291 in *Rasch Models: Foundations, Recent Developments, and Applications*, eds. G. Fischer and I. Molenaar. New York: Springer-Verlfag.

Anderson, J. A. 1972. Separate sample logistic discrimination. *Biometrika* 59: 19-35.

Anderson, J. A. 1975. Quadratic logistic discrimination. *Biometrika* 62: 149-154.

Anderson, J. A. 1984. Regression and ordered categorical variables. *J. Roy. Statist. Soc. Ser B* 46: 1-30.

Anderson, D. A., and M. Aitkin. 1985. Variance component models with binary response: Interviewer variability. *J. Roy. Statist. Soc. Ser B* 47: 203-210.

Anderson, C. J., and U. Böckenholt. 2000. Graphical regression models for polytomous variables. *Psychometrika* 65: 497-509.

Anderson, T. W., and L. A. Goodman. 1957. Statistical inference about Markov chains. *Ann. Math. Statist.* 28: 89-110.

Anderson, J. A., and P. R. Philips. 1981. Regression, discrimination, and measurement models for ordered categorical variables. *Appl. Statist.* 30: 22-31.

Anderson, C. J., and J. K. Vermunt. 2000. Log-multiplicative models as latent variable models for nominal and/or ordinal data. *Sociol. Methodol.* 30: 81-121.

Aranda-Ordaz, F. J. 1981. On two families of transformations to additivity for binary response data. *Biometrics* 68: 357-363.

Aranda-Ordaz, F. J. 1983. An extension of the proportional hazards model for grouped data. *Biometrics* 39: 109-117.

Arminger, G., C. C. Clogg, and T. Cheng. 2000. Regression analysis of multivariate binary response variables using Rasch-type models and finite mixture methods. *Sociol. Methodol.* 30: 1-26.

Armitage, P. 1955. Tests for linear trends in proportions and frequencies. *Biometrics* 11: 375-386.

Ashford, J. R., and R. D. Sowden. 1970. Multivariate probit analysis. *Biometrics* 26: 535-546.

Asmussen, S., and D. Edwards. 1983. Collapsibility and response variables in contingency tables. *Biometrika* 70: 567-578.

Azzalini, A. 1994. Logistic regression for autocorrelated data with application to repeated measures. *Biometrika* 81: 767-775.

Baglivo, J., D. Olivier, and M. Pagano. 1992. Methods for exact goodness-of-fit tests. *J. Amer. Statist. Assoc.* 87: 464-469,

Baker, S. G. 1992. A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *J. Comput. Graph. Statist.* 1: 63-76.

Baker, S. G., and N. M. Laird. 1988. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Amer. Statist. Assoc.* 83: 62-69.

Baker, R. J., M. R. B. Clarke, and P. W. Lane. 1985. Zero entries in contingency tables. *Comput. Statist. Data Anal.* 3: 33-45.

Banerjee, C., M. Capozzoli, L. McSweeney, and D. Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *Canad. J. Statist.* 27: 3-23.

Baptista, J, and M. C. Pike. 1977. Algorithm AS115: Exact two-sided confidence limits for the odds ratio in a 2 ×2 table. *Appl. Statist.* 26: 214-220.

Barnard, G. A. 1945. A new test for 2 ×2 tables. *Nature* 156: 177.

Barnard, G. A. 1947. Significance tests for 2 × 2 tables. *Biometrika* 34: 123-138.

Barnard, G. A. 1949. Statistical inference. *J. Roy. Statist. Soc. Ser B* 11: 115-139.

Barnard, G. A. 1979. In contradiction to J. Berkson's dispraise: Conditional tests can be more efficient. *J. Statist. Plann. Inference* 3: 181-188.

Barndorff-Nielsen, O. E., and B. Jörgensen. 1991. Some parametric models on the simplex. *J. Multivariate Anal.* 39: 106-116.

Bartholomew, D. J. 1980. Factor analysis for categorical data. *J. Roy. Statist. Soc. Ser B* 42: 293-321.

Bartholomew, D. J., and M. Knott. 1999. *Latent Variable Models and Factor Analysis*, 2nd ed. London: Edward Arnold.

Bartlett, M. S. 1935. Contingency table interactions. *J. Roy. Statist. Soc. Suppl.* 2: 248-252.

Bartlett, M. S. 1937. Some examples of statistical methods of research in agriculture and applied biology. *J. Roy.*

Statist. Soc. Suppl. 4: 137-183.

Becket, M. 1989a. Models for the analysis of association in multivariate contingency tables. *J. Amer. Statist. Assoc.* 84: 1014-1019.

Becket, M. 1989b. On the bivariate normal distribution and association models for ordinal categorical data. *Statist. Probab. Lett.* 8: 435-440.

Becker, M. 1990. Maximum likelihood estimation of the RC (M) association model. *Appl. Statist.* 39: 152-167.

Becker, M., and A. Agresti. 1992. Log-linear modelling of pairwise interobserver agreement on a categorical scale. *Statist. Medic.* 11: 101-114.

Becker, M., and C. C. Clogg. 1989. Analysis of sets of two-way contingency tables using association models. *J. Amer. Statist. Assoc.* 84: 142-151.

Bedrick, E. J. 1983. Chi-squared tests for cross-classified tables of survey data. *Biometrika* 70: 591-595.

Bedrick, E. J. 1987. A family of confidence intervals for the ratio of two binomial proportions. *Biometrics* 43: 993-998.

Begg, C. B., and R. Gray. 1984. Calculation of polytomous logistic regression parameters using individualized regressions. *Biometrika* 71: 11-18.

Beitler, P. J., and J. R. Landis. 1985. A mixed-effects model for categorical data. *Biometrics* 41: 991-1000.

Benedetti, J. K., and M. B. Brown. 1978. Strategies for the selection of loglinear models. *Biometrics* 34: 680-686.

Benichou, J. 1998. Attributable risk. Pp. 216-229 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Benzécri, J. -P. 1973. *L'Analyse des Données*, Vol. 1, *La Taxonomie*; Vol. 2, *L'Analyse des Correspondances*. Paris: Dunod.

Berger, R., and D. D. Boos. 1994. *p*-Values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* 89: 1012-1016.

Bergsma, W. P., and T. Rudas. 2002. Marginal models for categorical data. *Ann. Statist.* 30:140-159.

Berkson, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Amer. Statist. Assoc.* 33: 526-536.

Berkson, J. 1944. Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc.* 39:357-365.

Berkson, J. 1951. Why I prefer Logits to probits. *Biometrics* 7: 327-339.

Berkson, J. 1953. A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function. *J. Amer. Statist. Assoc.* 48: 565-599.

Berkson, J. 1955. Maximum likelihood and minimum Logit χ^2 estimation of the logistic function. *J. Amer. Statist. Assoc.* 50: 130-162.

Berkson, J. 1978. In dispraise of the exact test. *J. Statist. Plann. Inference* 2: 27-42.

Berkson, J. 1980. Minimum chi-square, not maximum likelihood! *Ann. Statist.* 8: 457-487.

Berry, G., and P. Armitage. 1995. Mid-*P* cofidence intervals: A brief review. *The Statistician* 44: 417-423.

Bhapkar, V. P. 1966. A note on the equivalence of two test criteria for hypotheses in categorical data. *J. Amer. Statist. Assoc.* 61: 228-235.

Bhapkar, V. P. 1968. On the analysis of contingency tables with a quantitative response. *Biometrics* 24: 329-338.

Bhapkar, V. P. 1973. On the comparison of proportions in matched samples. *Sankhya Ser A* 35: 341-356.

Bhapkar, V. P. 1989. Conditioning on ancillary statistics and loss of information in the presence of nuisance parameters. *J. Statist. Plann. Inference*. 21: 139-160.

Bhapkar, V. P., and G. G. Koch. 1968. On the hypothesis of "no interaction" in multidimensional contingency tables. *Biometrics* 24: 567-594.

Bhapkar, V. P., and G. W. Somes. 1977. Distribution of *Q* when testing equality of matched proportions. *J. Amer. Statist. Assoc.* 72: 658-661.

Biggeri, A. 1998. Negative binomial distribution. Pp. 2962-2967 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Billingsley, P. 1961. Statistical methods in Markov chains. *Ann. Math. Statist.* 32: 12-40.

Birch, M. W. 1963. Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. Ser. B* 25: 220-233.

Birch, M. W. 1964a. A new proof of the Pearson-Fisher theorem. *Ann. Math. Statist.* 35: 817-824.

Birch, M. W. 1964b. The detection of partial association Ⅰ: The 2 × 2 case. *J. Roy. Statist. Soc. Ser. B* 26: 313-324.

Birch, M. W. 1965. The detection of partial association Ⅱ: The general case. *J. Roy. Statist. Soc. Ser B* 27: 111-124.

Bishop, Y. M. M. 1971. Effects of collapsing multidimensional contingency tables. *Biometrics* 27: 545-562.

Bishop, Y. M. M., and F. Mosteller. 1969. Smoothed contingency table analysis. Chap. Ⅳ-3 in *The National Halothane Study*. Washington, DC: U. S. Government Printing Office.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

Blaker, H. 2000. Confidence curves and improved exact confidence intervals for discrete distributions. *Canad. J. Statist.* 28: 783-798.

Bliss, C. I. 1934. The method of probits. *Science* 79: 38-39.

Bliss, C. I. 1935. The calculation of the dosage-mortality curve. *Ann. Appl. Biol.* 22: 134-167.

Blyth, C. R. 1972. On Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Assoc.* 67: 364-366.

Blyth, C. R., and H. A. Still. 1983 Binomial confidence intervals. *J. Amer. Statist. Assoc.* 78: 108-116.

Bock, R. D. 1970. Estimating multinomial response relations. Pp. 453-479 in *Contributions to Statistics and Probability*, ed. R. C. Bose. Chapel Hill, NC: University of North Carolina Press.

Bock, R. D., and M. Aitkin. 1981. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 46: 443-459.

Bock, R. D., and L. V. Jones. 1968. *The Measurement and Prediction of Judgement and Choice*. San Francisco: Holden-Day.

Böckenholt, U., and W. Dillon. 1997. Modelling within-subject dependencies in ordinal paired comparison data. *Psychometrika* 62: 411-434.

Bonney, G. E. 1987. Logistic regression for dependent binary observations. *Biometrics* 43: 951-973.

Boos, D. D. 1992. On generalized score tests. *Amer. Statist.* 46: 327-333.

Booth, J., and R. Butler. 1999. An importance sampling algorithm for exact conditional tests in log-linear models. *Biometrika* 86: 321-332.

Booth, J. G., and J. P. Hobert. 1998. Standard errors of prediction in generalized linear mixed models. *J. Amer. Statist. Assoc.* 93: 262-272.

Booth, J. G., and J. P. Hobert. 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. Roy. Statist. Soc. Ser. B* 61: 265-285.

Bowker, A. H. 1948. A test for symmetry in contingency tables. *J. Amer. Statist. Assoc.* 43: 572-574.

Box, J. F. 1978. *R. A. Fisher: The Life of a Scientist.* New York: Wiley.

Bradley, R. A. 1976. Science, statistics, and paired comparisons. *Biometrics* 32: 213-240.

Bradley, R. A., and M. E. Terry. 1952. Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika* 39: 324-345.

Breslow, N. 1976. Regression analysis of the log odds ratio: A method for retrospective studies. *Biometrics* 32: 409-416.

Breslow, N. 1981. Odds ratio estimators when the data are sparse. *Biometrika* 68: 73-84.

Breslow, N. 1982. Covariance adjustment of relative-risk estimates in matched studies. *Biometrics* 38: 661-672.

Breslow, N. 1984. Extra-Poisson variation in log-linear models. *Appl. Statist.* 33: 38-44.

Breslow, N. 1996. Statistics in epidemiology: The case-control study. *J. Amer. Statist. Assoc.* 91: 14-28.

Breslow, N., and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* 88: 9-25.

Breslow, N., and N. E. Day. 1980, 1987. *Statistical Methods in Cancer Research*, Vol. I, *The Analysis of Case-Control Studies*; Vol. Ⅱ. *The Design and Analysis of Cohort Studies*. Lyon: IARC.

Breslow, N., and X. Lin. 1995. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* 82: 81-91.

Breslow, N., and W. Powers. 1978. Are there two logistic regressions for retrospective studies? *Biometrics* 34: 100-105.

Breslow, N., N. Day, K. Halvorsen, R. Prentice, and C. Sabaí. 1978. Estimation of multiple relative risk functions in matched case-control studies. *Amer. J. Epidemiol.* 108: 299-307.

Brier, S. S. 1980. Analysis of contingency tables under cluster sampling. *Biometrika* 67: 591-596.

Brooks, S. P., B. J. T. Morgan, M. S. Ridout, and S. E. Pack. 1997. Finite mixture models for proportions. *Biometrics* 53: 1097-1115.

Bross, I. D. J. 1958. How to use ridit analysis. *Biometrics* 14: 18-38.

Brown, M. B. 1976. Screening effects in multidimensional contingency tables. *Appl. Statist.* 25: 37-46.

Brown, M. B., and J. K. Benedetti. 1977. Sampling behavior of tests for correlation in two-way contingency tables. *J. Amer. Statist. Assoc.* 72: 309-315.

Brown, P. J., and P. W. K. Rundell. 1985. Kernel estimates for categorical data. *Technometrics* 27: 293-299.

Brown, L. D., T. T. Cai, and A. Das Gupta. 2001. Interval estimation for a binomial proportion. *Statist. Sci.* 16: 101-133.

Brownstone, D., and K. F. Train. 1999. Forecasting new product penetration with flexible substitution patterns. *J. Econ-ometrics* 89: 109-129.

Bull, S. B., and A. Donner. 1987. The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. *J. Amer. Statist. Assoc.* 82: 1118-1122.

Burnham, K. P., and D. R. Anderson. 1998. *Model Selection and Inference: A Practical Information-Theoretic Approach.* New York: Springer-Verlag.

Burnham, K. P. and W. S. Overton. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 65: 625-633.

Burridge, J. 1981. A note on maximum likelihood estimation for regression models using grouped data. *J. Roy. Statist. Soc. Ser. B* 43: 41-45.

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data.* Cambridge, U. K. : Cambridge University Press.

Carey, V., S. L. Zeger, and P. Diggle. 1993. Modelling multivariate binary data with alternating logistic regressions. *Biometrika* 80: 517-526.

Carroll, R. J., S. Wang, and C. Y. Wang. 1995. Prospective analysis of logistic case-control pairs. *J. Amer. Statist. Assoc.* 90: 157-169.

Casella, G., and R. Berger. 2001. *Statistical Inference*, 2nd ed. Pacific Grove, CA: Wadsworth.

Catalano, P. J., and L. M. Ryan. 1992. Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Amer. Statist. Assoc.* 87: 651-658.

Caussinus, H. 1966. Contribution à l'analyse statistique des tableaux de corrélation. *Ann. Fac. Sci. Univ. Toulouse* 29: 77-182.

Chaloner, K., and K. Larntz. 1989. Optimal Bayesian design applied to logistic regression experiments. *J. Statist. Plann. Inference* 21: 191-208.

Chamberlain, G. 1980, Analysis of covariance with qualitative data. *Rev. Econ. Stud.* 47: 225-238.

Chambers, E. A., and D. R. Cox. 1967. Discrimination between alternative binary response models. *Biometrika* 54: 573-578.

Chambers, R. L., and D. G. Steel. 2001. Simple methods for ecological inference in 2 × 2 tables. *J. Roy. Statist. Soc. Ser. A* 164: 175-192.

Chan, I. 1998. Exact tests of equivalence and efficacy with non-zero lower bound for comparative studies. *Statist. Medic.* 17: 1403-1413.

Chan, J. S. K., and A. Y. C. Kuk. 1997. Maximum likelihood estimation for probit-linear mixed models with

correlated random effects. *Biometrics* 53: 86-97.

Chao, A., P. K. Tsay, S.-H. Lin, W.-Y. Shau, and D.-Y. Chao. 2001. The applications of capture-recapture models to epidemiological data. *Statist. Medic.* 20: 3123-3157.

Chapman, D. G., and R. C. Meng. 1966. The power of chi-square tests for contingency tables. *J. Amer. Statist. Assoc.* 61: 965-975.

Chen, Z. and L. Kuo. 2001. A note on the estimation of the multinomial Logit model with random effects. *Amer. Statist.* 55: 89-95.

Christensen, R. 1997. *Log-Linear Models and Logistic Regression*. New York: Springer-Verlag.

Chuang, C., D. Gheva, and C. Odoroff. 1985. Methods for diagnosing multiplicative-interaction models for two-way contingency tables. *Commun. Statist. Ser. A* 14: 2057-2080.

Clogg, C. C. 1995. Latent class models. Pp. 311-359 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, ed. G. Arminger and C. C. Clogg. New York: Plenum Press.

Clogg, C. C., and S. R. Eliason. 1987. Some common problems in log-linear analysis. *Sociol. Methods Res.* 15: 4-44.

Clogg, C. C., and L. A. Goodman. 1984. Latent structure analysis of a set of multidimensional contingency tables. *J. Amer. Statist. Assoc.* 79: 762-771.

Clogg, C. C., and E. S. Shihadeh. 1994. *Statistical Models for Ordinal Variables*. Thousand Oaks, CA: Sage Publications.

Clopper, C. J., and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26: 404-413.

Cochran, W. G. 1940. The analysis of variance when experimental errors follow the Poisson or binomial laws. *Ann. Math. Statist.* 11: 335-347.

Cochran, W. G. 1943. Analysis of variance for percentages based on unequal numbers. *J. Amer. Statist. Assoc.* 38: 287-301.

Cochran, W. G. 1950. The comparison of percentages in matched samples. *Biometrika* 37: 256-266.

Cochran, W. G. 1952. The χ^2 test of goodness-of-fit. *Ann. Math. Statist.* 23: 315-345.

Cochran, W. G. 1954. Some methods of strengthening the common χ^2 *tests. Biometrics* 10: 417-451.

Cochran, W. G. 1955. A test of a linear function of the deviations between observed and expected numbers. *J. Amer. Statist. Assoc.* 50: 377-397.

Coe, P. R., and A. C. Tamhane. 1993. Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Commun. Statist. Ser. B* 22: 925-938.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20: 37-46.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* 70: 213-220.

Cohen, A., and H. B. Sackrowitz. 1991. Tests for independence in contingency tables with ordered alternatives. *J. Multivariate Anal.* 36: 56-67.

Cohen, A., and H. B. Sackrowitz. 1992. An evaluation of some tests of trend in contingency tables. *J. Amer. Statist. Assoc.* 87: 470-475.

Collett, D. 1991. *Modelling Binary Data*. London: Chapman & Hall.

Conaway, M. R. 1989. Analysis of repeated categorical measurements with conditional likelihood methods. *J. Amer. Statist. Assoc.* 84: 53-62.

Cook, R. D., and S. Weisberg. 1999. *Applied Regression Including Computing and Graphics*. New York: Wiley.

Copas, J. B. 1973. Randomization models for the matched and unmatched 2 × 2 tables. *Biometrika* 60: 467-476.

Copas, J. B. 1983. Plotting *p* against *x*. *Appl. Statist.* 32: 25-31.

Copas, J. B. 1988. Binary regression models for contaminated data. *J. Roy. Statist. Soc. Ser B* 50: 225-265.

Corcoran, C., L. Ryan, P. Senchaudhuri, C. Mehta, N. Patel, and G. Molenberghs. 2001. An exact trend test for correlated binary data. *Biometrics* 57: 941-948.

Cormack, R. M. 1989. Log-linear models for capture-recapture. *Biometrics* 45: 395-413.

Cornfield, J. 1951. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast and cervix. *J. Natl. Cancer Inst.* 11: 1269-1275.

Cornfield, J. 1956. A statistical problem arising from retrospective studies. In *Proc. 3rd Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, 4: 135-148.

Cornfield, J. 1962. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis. *Fed. Proc.* 21, *Suppl.* 11: 58-61.

Coull, B. A., and A. Agresti. 1999. The use of mixed Logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* 55: 294-301.

Coull, B. A., and A. Agresti. 2000. Random effects modeling of multiple binomial responses using the multivariate binomial Logit-normal distribution. *Biometrics* 56: 73-80.

Cox, C. 1984. An elementary introduction to maximum likelihood estimation for multinomial models: Birch's theorem and the delta method. *Amer. Statist.* 38: 283-287.

Cox, C. 1995. Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statist. Medic.* 14: 1191-1203.

Cox, C. 1996. Nonlinear quasi-likelihood models: Applications to continuous proportions. *Comput. Statist. Data Anal.* 21: 449-461.

Cox, D. R. 1958a. The regression analysis of binary sequences. *J. Roy. Statist. Soc. Ser. B* 20: 215-242.

Cox, D. R. 1958b. Two further applications of a model for binary regression. *Biometrika* 45: 562-565.

Cox, D. R. 1970. *The Analysis of Binary Data* (2nd ed. 1989, by D. R. Cox and E. J. Snell). London: Chapman & Hall.

Cox, D. R. 1972. The analysis of multivariate binary data. *Appl. Statist.* 21: 113-120.

Cox, D. R. 1983. Some remarks on overdispersion. *Biometrika* 70: 269-274.

Cox, D. R., and D. V. Hinkley. 1974. *Theoretical Statistics*. London: Chapman & Hall.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Cressie, N., and T. R. C. Read. 1984. Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* 46: 440-464.

Cressie, N., and T. R. C. Read. 1989. Pearson X^2 and the loglikelihood ratio statistic G^2: A comparative review. *Internat. Statist. Rev.* 57: 19-43.

Croon, M., W. Bergsma, and J. Hagenaars. 2000. Analyzing change in categorical variables by generalized log-linear models. *Sociol. Methods Res.* 29: 195-229.

Crouchley, R. 1995. A random-effects model for ordered categorical data. *J. Amer. Statist. Assoc.* 90: 489-498.

Crowder, M. J. 1978. Beta-binomial ANOVA for proportions. *Appl. Statist.* 27: 34-37.

D'Agostino, R. B., Jr. 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statist. Medic.* 17: 2265-2281.

Daniels, M. J., and C. Gatsonis. 1999. Hierarchical generalized linear models in the analysis of variations in health care utilization. *J. Amer. Statist. Assoc.* 94: 29-42.

Dardanoni, V., and A. Forcina. 1998. A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *J. Amer. Statist. Assoc.* 93: 1112-1123.

Darroch, J. N. 1962. Interactions in multi-factor contingency tables. *J. Roy. Statist. Soc. Ser. B* 24: 251-263.

Darroch, J. N. 1981. The Mantel-Haenszel test and tests of marginal symmetry; Fixed-effects and mixed models for a categorical response. *Intemat. Statist. Rev.* 49: 285-307.

Darroch, J. N., and P. I. McCloud. 1986. Category distinguishability and observer agreement. *Austral. J. Statist.* 28: 371-388.

Darroch, J. N., and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* 43: 1470-1480.

Darroch, J. N., S. L. Lauritzen, and T. P. Speed. 1980. Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* 8: 522-539.

Darroch, J. N., S. E. Fienherg, G. F. V. Glonek, and B. W. Junker. 1993. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *J. Amer. Statist. Assoc.* 88: 1137-1148.

Das Gupta, S., and M. D. Perlman. 1974. Power of the noncentral F-test: Effect of additional variates on Hotelling's T^2-test. *J. Amer. Statist. Assoc.* 69: 174-180.

David, H. A. 1988. *The Method of Paired Comparisons*, 2nd ed. Oxford: Oxford University Press.

Davis, L. J. 1986a. Exact tests for 2 by 2 contingency tables. *Amer. Statist.* 40: 139-141.

Davis, L. J. 1986b. Relationship between strictly collapsible and perfect tables. *Statist. Probab. Lett.* 4: 119-122.

Davis, L. J. 1989. Intersection union tests for strictly collapsibility in three-dimensional contingency tables. *Ann. Statist.* 17: 1693-1708.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge, U. K. Cambridge University Press.

Dawson, R. B., Jr. 1954. A simplified expression for the variance of the χ^2-function on a contingency table. *Biornetrika* 41: 280.

Day, N. E., and D. P. Byar. 1979. Testing hypotheses in case-control studies: Equivalence of Mantel-Haenszel statistics and Logit score tests. *Biometrics* 35: 623-630.

de Falguerolles, A., S. Jmel, and J. Whittaker. 1995. Correspondence analysis and association models constrained by a conditional independence graph. *Psychometrika* 60: 161-180.

Deming, W. E. 1964. *Statistical Adjustment of Data* (reprint of 1943 Wiley text). New York: Dover.

Deming, W. E., and F. F. Stephan. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* 11: 427-444.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 39: 1-38.

Dey, D. K., S. K. Ghosh, and B. K. Mallick (editors). 2000. *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.

Diaconis, P., and B. Efron. 1985. Testing for independence in a two-way table: New interpretations of the chi-square statistic. *Ann. Statist.* 13: 845-874.

Diaconis, P., and B. Sturmfels. 1998. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* 26: 363-397.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger. 2002. *Analysis of Longitudinal Data*, 2nd ed. Oxford: Clarendon Press.

Dittrich, R., R. Hatzinger, and W. Katzenbeisser. 1998. Modeling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Appl. Statist.* 47: 511-525.

Dobson, A. J. 2001. *An Introduction to Generalized Linear Models, 2nd ed.* London: Chapman & Hall.

Dong, J. 1998. Simpson's paradox. Pp. 4108-4110 in *Encyclopedia of Biostatistics*, Vol. 5. Chichester, UK: Wiley.

Dong, J., and J. S. Simonoff. 1994. The construction and properties of boundary kernels for smoothing sparse multinomials. *J. Computat. Graph. Statist.* 3:57-66.

Dong, J., and J. S. Simonoff. 1995. A geometric combination estimator for d-dimensional ordinal sparse contingency tables. *Ann. Statist.* 23:1143-1159.

Donner, A., and W. W. Hauck. 1986. The large-sample efficiency of the Mantel-Haenszel estimator in the fixed-strata case. *Biometrics* 42: 537-545.

Doolittle, M. H. 1888. Association ratios. *Bull. Philos. Soc. Washington* 10: 83-87, 94-96.

Drost, F. C., W. C. M. Kallenberg, D. S. Moore, and J. Oosterhoff. 1989. Power approximations to multinomial tests of fit. *J. Amer. Statist. Assoc.* 84: 130-141.

Ducharme, G. R., and Y. Lepage. 1986. Testing

collapslbfitty in contingency tables. *J. Roy. Statist. Soc. Ser B* 48: 197-205.

Dupont, W. D. 1986. Sensitivity of Fisher's exact test to minor perturbations in 2 × 2 contingency tables. *Statist. Medic.* 5: 629-635.

Dyke, G. V., and H. D. Patterson. 1952. Analysis of factorial arrangements when the data are proportions. *Biometrics* 8: 1-12.

Edwardes, M. D. deB. 1997. Univariate random cut-points theory for the analysis of ordered categorical data. *J. Amer. Statist. Assoc.* 92:1114-1123.

Edwards, A. W. F. 1963. The measure of association in a 2 × 2 table. *J. Roy. Statist. Soc. Ser A* 126: 109-114.

Edwards, D. 2000. *Introduction to Graphical Modelling*, 2nd ed. New York: Springer-Verlag.

Edwards, D., and S. Kreiner. 1983. The analysis of contingency tables by graphical models. *Biometrika* 70: 553-565.

Efron, B. 1975. The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.* 70: 892-898.

Efron, B. 1978. Regression and ANOVA with zero-one data: Measures of residual variation. *J. Amer. Statist. Soc.* 73: 113-121.

Efron, B., and D. V. Hinkley. 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* 65:457-482.

Efron, B., and C. Morris. 1975. Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* 70: 311-319.

Ekholm, A., J. W. McDonald, and P. W. F. Smith. 2000. Association models for a multivariate binary response. *Biometrics* 56: 712-718.

Escoufier, Y. 1982. L'analyse des tableaux de contingence simples et multiples. In *Proc. International Meeting on the Analysis of Multidimensional Contingency Tables* (Rome, 1981), ed. R. Coppi. *Metron* 40: 53-77.

Espeland, M. A., and S. L. Handelman. 1989. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* 45: 587-599.

Fahrmeir, L., and G. Tutz. 2001. *Multiuariate Statistical Modelling based on Generalized Linear Models*, 2nd ed. New York: Springer-Verlag.

Farewell, V. T. 1979. Some results on the estimation of logistic models based on retrospective data. *Biometrika* 66: 27-32.

Farewell, V. T. 1982. A note on regression analysis of ordinal data with variability of classification. *Biometrika* 69: 533-538.

Fay, R. 1985. A jackknifed chi-squared test for complex samples. *J. Amer. Statist. Assoc.* 80: 148-157.

Fay, R. 1986. Causal models for patterns of nonresponse. *J. Amer. Statist. Assoc.* 81: 354-365.

Ferguson, T. S. 1967. *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.

Fienberg, S. E. 1970a. An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.* 41: 907-917.

Fienberg, S. E. 1970b. Quasi-independence and maximum likelihood estimation in incomplete contingency tables. *J. Amer. Statist. Soc.* 65: 1610-1616.

Fienberg, S. E. 1972. The analysis of incomplete multi-way contingency tables. *Biometrics* 28: 177-202.

Fienberg, S. E. 1980. Fisher's contributions to the analysis of categorical data. Pp. 75-84 in *R. A. Fisher: An Appreciation*, ed. S. E. Fienberg and D. V. Hinkley. Berlin: Springer-Verlag.

Fienberg, S. E. 1984. The contributions of William Cochran to categorical data analysis. Pp. 103-118 in *W. G. Cochran's Impact on Statistics*, ed. P. S. R. S. Rao and J. Sedransk. New York: Wiley.

Fienberg, S. E., and P. W. Holland. 1973. Simultaneous estimation of multinomial cell probabilities. *J. Amer. Statist. Assoc.* 68: 683-690.

Fienberg, S. E., and K. Larntz. 1976. Loglinear representation for paired and multiple comparison models. *Biometrika* 63: 245-254.

Fienberg, S. E., M. A. Johnson, and B. J. Junker. 1999. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *J. Roy. Statist. Soc. Ser. A* 162: 383-405.

Finney, D. J. 1947. The estimation from individual records of the relationship between dose and quantal response. *Biometrika* 34: 320-334.

Finney, D. J. 1971. *Probit Analysis*, 3rd ed. Cambridge: Cambridge University Press.

Firth, D. 1987. On the efficiency of quasi-likelihood estimation. *Biometrika* 74: 233-245.

Firth, D. 1989. Marginal homogeneity and the superposition of Latin squares. *Biomettika* 76: 179-182.

Firth, D. 1991. Generalized linear models. Pp. 55-82 in *Statistical Theory and Modelling. In Honour of Sir David Cox, FRS*, D. V. Hinkley, N. Reid, and E. J. Snell, eds. London: Chapman & Hall.

Firth, D. 1993a. Bias reduction of maximum likelihood estimates. *Biomettika* 80: 27-38.

Firth, D. 1993b. Recent developments in quasi-likelihood methods. *Proc. ISI 49th Session*, pp. 341-358.

Firth, D., and J. Kuha. 2000. On the index of dissimilarity for lack of fit in log linear models. Unpublished manuscript.

Fischer, G. H., and I. W. Molenaar. 1995. *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer-Verlag.

Fisher, R. A. 1922. On the interpretation of chi-square from contingency tables, and thc calculation of P. *J. Roy. Statist. Soc.* 85: 87-94.

Fisher, R. A. 1924. The conditions under which chi-square measures the discrepancy between observation and hypothesis. *J. Roy. Statist. Soc.* 87: 442-450.

Fisher, R. A. 1926. Bayes' theorem and the fourfold table. *Eugenics Rev.* 18:32-33.

Fisher, R. A. 1934, 1970. *Statistical Methods for Research Workers* (originally published 1925, 14th ed., 1970.) Edinburgh: Oliver & Boyd.

Fisher, R. A. 1935a. *The Design of Experiments* (8th ed., 1966). Edinburgh: Oliver & Boyd.

Fisher, R. A. 1935b. Appendix to article by C. Bliss. *Ann. Appl. Biol.* 22:164-165.

Fisher, R. A. 1935c. The logic of inductive inference. *J. Roy. Statist. Soc.* 98: 39-82.

Fisher, R. A. 1945. A new test for 2 × 2 tables (Letter to the Editor). *Nature* 156: 388.

Fisher, R. A. 1956. *Statistical Methods for Scientific Inference*. Edinburgh: Oliver & Boyd.

Fisher, R. A., and F. Yates. 1938. *Statistical Tables*. Edinburgh: Oliver and Boyd.

Fitzmaurice, G. M., and N. M. Laird. 1993. A likelihood-based method for analysing longitudinal binary responses. *Biometrika* 80: 141-151.

Fitzmaurice, G. M., N. M. Laird, and S. Lipsitz. 1994. Analysing incomplete longitudinal binary responses: A likelihood-based approach. *Biometrics* 50: 601-612.

Fitzmaurice, G. M., N. M. Laird, and A. G. Rotnitzky. 1993. Regression models for discrete longitudinal responses. *Statist. Sci.* 8: 284-299.

Fitzpatrick, S., and A. Scott. 1987. Quick simultaneous confidence intervals for multinomial proportions. *J. Amer. Statist. Assoc.* 82: 875-878.

Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions*, 2nd ed. New York: Wiley.

Fleiss, J. L., and J. Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Paychol. Meas.* 33: 613-619.

Fleiss, J. L., J. Cohen, and B. S. Everitt. 1969. Large-sample standard errors of kappa and weighted kappa. *Psychol. Bull.* 72: 323-327.

Follman, D. A., and D. Lambert. 1989. Generalizing logistic regression by nonparametric mixing. *J. Amer. Statist. Assoc.* 84: 295-300.

Forster, J. J., and P. W. F. Smith. 1998. Model-based inference for categorical survey data subject to non-ignorable non-response. *J. Roy. Statist. Soc. Ser B* 60: 57-70.

Forster, J. J., J. W. McDonald, and P. W. F. Smith. 1996. Monte Carlo exact conditional tests for log-linear and logistic models, *J. Roy. Statist. Soc. Ser B* 58: 445-453.

Fowlkes, E. B. 1987. Some diagnostics for binary logistic regression via smoothing. *Biometrika* 74: 503-515.

Fowlkes, E. B., A. E. Freeny, and J. Landwehr. 1988. Evaluating logistic models for large contingency tables. *J. Amer. Statist. Assoc.* 83: 611-622.

Freedman, D., R. Pisani, and R. Purves. 1978. *Statistics*. New York: W. W. Norton.

Freeman, G. H., and J. H. Halton. 1951. Note on an exact treatment of contingency, goodness- of-fit and other problems of significance. *Biometrika* 38: 141-149.

Freeman, D. H., Jr. and T. R. Holford. 1980. Summary rates. *Biometrics* 36: 195-205.

Freeman, M. F., and J. W. Tukey. 1950. Transformations related to the angular and the square root. *Ann. Math. Statist.* 21: 607-611.

Freidlin, B., and J. L. Gastwirth. 1999. Unconditional versions of several tests commonly used in the analysis of contingency tables. *Biometrics* 55: 264-267.

Friendly, M. 2000. *Visualizing Categorical Data*. Cary, NC: SAS Institute.

Frome, E. L. 1983. The analysis of rates using Poisson regression models. *Biometrics* 39: 665-674.

Fuchs, C. 1982. Maximum likelihood estimation and model selection in contingency tables with missing data. *J. Amer. Statist. Assoc.* 77: 270-278.

Gabriel, K. R. 1966. Simultaneous test procedures for multiple comparisons on categorical data. *J. Amer. Statist. Assoc.* 61: 1081-1096.

Gabriel, K. R. 1971. The biplot graphic display of matrices with applications to principal component analysis. *Biometrika* 58: 453-467.

Gail, M. H., and J. J. Gart. 1973. The determination of sample sizes for use with the exact conditional test in 2 ×2 comparative trials. *Biometrics* 29: 441-448.

Gall, M., and N. Mantel. 1977. Counting the number of $r \times c$ contingency tables with fixed margins. *J. Amer. Statist. Assoc.* 72: 859-862.

Gart, J. J. 1966. Alternative analyses of contingency tables. *J. Roy. Statist. Soc. Ser B* 28:164-179.

Gart, J. J. 1969. An exact test for comparing matched proportions in crossover designs. *Biometrika* 56: 75-80.

Gatt, J. J. 1970. Point and interval estimation of the common odds ratio in the combination of 2 × 2 tables with fixed margins. *Biometrika* 57: 471-475.

Gart, J. J. 1971. The comparison of proportions: A review of significance tests, confidence intervals and adjustments for stratification. *Rev. lnternat. Statist. Rev.* 39: 148-169.

Gart, J. J., and J. Nam. 1988. Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* 44: 323-338.

Gart, J. J., and J. R. Zweifel. 1967. On the bias of various estimators of the Logit and its variance with appliations to quantal bioassay. *Biometrika* 54: 181-187.

Gelfand, A. E., and A. F. Smith. 1990. Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* 85: 398-409.

Genter, F. C., and V. T. Farewell. 1985. Goodness-of-link testing in ordinal regression models. *Canad. J. Statist.* 13: 37-44.

Ghosh, B. K. 1979. A comparison of some approximate confidence intervals for the binomial parameter. *J. Amer. Statist. Assoc.* 74: 894-900.

Ghosh, M., M. Chen, A. Ghosh, and A. Agresti. 2000. Hierarchical Bayesian analysis of binary matched pairs data. *Statist. Sin.* 10: 647-657.

Gibbons, R. D., and D. Hedeker. 1997. Random-effects probit and logistic regression models for three-level data. *Biometrics* 53: 1527-1537.

Gill, J. 2000. *Generalized Linear Models: A Unified Approach*. Thousand Oaks, CA: Sage Publications.

Gilmour, A. R., R. D. Anderson, and A. L. Rae. 1985. The analysis of binomial data by a generalized linear mixed model. *Biometrika* 72: 593-599.

Gilula, Z., and S. Haberman. 1986. Canonical analysis of contingency tables by maximum likelihood. *J. Amer. Statist. Assoc.* 81: 780-788.

Gilula, Z., and S. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *J. Amer. Statist.*

Assoc. 83: 760-771.

Gilula, Z., and S. Haberman. 1998. Chi-square, partition of. Pp. 622- 627 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Gleser, L. J, and D. S. Moore. 1985. The effect of positive dependence on chi-squared tests for categorical data. *J. Roy. Statist. Soc. Ser B* 47: 459-465.

Glonek, G. 1996. A class of regression models for multivariate categorical responses. *Biometrika* 83: 15-28.

Glonek, G. F. V., and P. McCullagh. 1995. Multivariate logistic models. *J. Roy. Statist. Soc. Ser.* B 57: 533-546.

Glonek, G., J. N. Darroch, and T. P. Speed. 1988. On the existence of maximum likelihood estimators for hierarchical loglinear models. *Scand. J. Statist.* 15: 187-193.

Gokhale, D. V., and S. Kullback. 1978. *The Information in Contingency Tables*. New York: Marcel Dekker.

Goldstein, H. 1995. *Multilevel Statistical Models*, 2nd ed. London: Edward Arnold.

Goldstein, H., and J. Rasbash. 1996. Improved approximations for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser A* 159: 505-513.

Good, I. J. 1963. Maximum entropy for hypothesis formulation, especially for multi-dimensional contingency tables. *Ann. Math. Statist.* 34: 911-934.

Good, I. J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

Good, I. J. 1976. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* 4: 1159-1189.

Good, I. J., and R. A. Gaskins. 1971. Nonparametric roughness penalties for probability densities. *Biometrika* 58: 255-277.

Good, I. J., and Y. Mittal. 1987. The amalgamation and geometry of two-by-two contingency tables. *Ann. Statist.* 15: 694-711.

Good, I. J., T. N. Gover, and G. J. Mitchell. 1970. Exact distributions for χ^2 and for the likelihood-ratio statistic for the equiprobable multinomial distribution. *J. Amer. Statist. Assoc.* 65: 267-283.

Goodman, L. A. 1964a. Simultaneous confidence intervals for cross-product ratios in contingency tables. *J. Roy. Statist. Soc. Ser B* 26: 86-102.

Goodman, L. A. 1964b. Interactions in multi-dimensional contingency tables. *Ann. Math. Statist.* 35: 632-646.

Goodman, L. A. 1965. On simultaneous confidence intervals for multinomial proportions. *Technometrics* 7: 247-254.

Goodman, L. A. 1968. The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. *J. Amer. Statist. Assoc.* 63: 1091-1131.

Goodman, L. A. 1969a. On partitioning chi-square and detecting partial association in three-way contingency tables. *J. Roy. Statist. Soc. Ser B* 31: 486-498.

Goodman, L. A. 1969b. How to ransack social mobility tables and other kinds of cross-classification tables. *Amer. J. Sociol.* 75: 1-40.

Goodman, L. A. 1970. The multivariate analysis of qualitative data: Interaction among multiple classifications. *J. Amer. Statist. Assoc.* 65: 226-256.

Goodman, L. A. 1971a. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* 13: 33-61.

Goodman, L. A. 1971b. The partitioning of chi-square, the analysis of marginal contingency tables, and the estimation of expected frequencies in multidimensional contingency tables. *J. Amer. Statist. Assoc.* 66: 339-344.

Goodman, L. A. 1973. The analysis of multidimensional contingency tables with some variables are posterior to others: A modified path analysis approach. *Biometrika* 60: 179-192.

Goodman, L. A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61: 215-231.

Goodman, L. A. 1979a. Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* 74: 537-552.

Goodman, L. A. 1979b. Multiplicative models for square contingency tables with ordered categories. *Biometrika* 66: 413-418.

Goodman, L. A. 1981a. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* 76: 320-334.

Goodman, L. A. 1981b. Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika* 68: 347-355.

Goodman, L. A. 1983. The analysis of dependence in cross-classification having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics* 39: 149-160.

Goodman, L. A. 1985. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.* 13: 10-69.

Goodman, L A. 1986. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *Internat. Statist. Rev.* 54: 243-309.

Goodman, L. A. 1996. A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. *J. Amer. Statist. Assoc.* 91: 408-427.

Goodman, L. A. 2000. The analysis of cross-classified data: Notes on a century of progress in contingency table analysis, and some comments on its prehistory and its future. Pp. 189-231 in *Statistics for the 21st Century*, ed. C. R. Rao and G. J. Székely. New York: Marcel Dekker.

Goodman, L. A., and W. H. Kruskal. 1979. *Measures of Association for Cross Classifications*. New York: Springer-Verlag (contains articles appearing in J.

Amer. Statist. Assoc. in 1954, 1959, 1963, 1972).

Gould, S. J. 1981. *The Mismeasure of Man*. New York: W. W. Norton.

Gourieroux, C., A. Monfort, and A. Trognon. 1984. Pseudo maximum likelihood methods: Theory. *Econometrica* 52: 681-700.

Graubard, B. I., and E. L. Korn. 1987. Choice of column scores for testing independence in ordered 2 × *K* contingency tables. *Biometrics* 43: 471-476.

Green, P. J. 1984. Iteratively weighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *J. Roy. Statist. Soc. Ser B* 46: 149-192.

Greenacre, M. J. 1993. *Correspondence Analysis in Practice*. New York: Academic Press.

Greenland, S. 1991. On the logical justification of conditional tests for two-by-two contingency tables. *Amer. Statist.* 45: 248-251.

Greenland, S, and J. M. Robins. 1985. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* 41: 55-68.

Greenwood, M., and G. U. Yule. 1920. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Statist. Soc. Ser A* 83: 255-279.

Greenwood, P. E., and M. S. Nikulin. 1996. *A Guide to Chi-Squared Testing*. New York: Wiley.

Grizzle, J. E., C. F. Starmer, and G. G. Koch. 1969. Analysis of categorical data by linear models. *Biometrics* 25: 489-504.

Gross, S. T. 1981. On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications. *J. Amer. Statist. Assoc.* 76: 935-941.

Gueorguieva, R., and A. Agresti. 2001. A correlated probit model for joint modeling of clustered binary and continuous responses. *J. Amer. Statist. Assoc.* 96: 1102-1112.

Haber, M. 1980. A comparison of some continuity corrections for the chi-squared test on 2 × 2 tables. *J. Amer. Statist. Assoc.* 75: 510-515.

Haber, M. 1982. The continuity correction and statistical testing. *Internat. Statist. Rev.* 50: 135-144.

Haber, M. 1985. Maximum likelihood methods for linear and log-linear models in categorical data. *Comput. Statist. Data Anal.* 3: 1-10.

Haber, M. 1986. An exact unconditional test for the 2 × 2 comparative trial. *Psychol. Bull.* 99: 129-132.

Haber, M. 1989. Do the marginal totals of a 2 × 2 contingency table contain information regarding the table proportions? *Commun. Statist. Ser A* 18: 147-156.

Haberman, S. J. 1973a. The analysis of residuals in cross-classification tables. *Biometrics* 29: 205-220.

Haberman, S. J. 1973b. Log-linear models for frequency data: Sufficient statistics and likelihood equations. *Ann. Statist.* 1:617-632.

Haberman, S. J. 1974a. *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Haberman, S. J. 1974b. Log-linear models for frequency tables with ordered classifications. *Biometrics* 36: 589-600.

Haberman, S. J. 1977a. Log-linear models and frequency tables with small expected cell counts. *Ann. Statist.* 5: 1148-1169.

Haberman, S. J. 1977b. Maximum likelihood estimation in exponential response models. *Ann. Statist.* 5: 815-841.

Haberman, S. J. 1978, 1979. *Analysis of Qualitative Data*, Vols. 1 and 2. New York: Academic Press.

Haberman, S. J. 1981. Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. *Ann. Statist.* 9: 1178-1186.

Haberman, S. J. 1982. The analysis of dispersion of multinomial responses. *J. Amer. Statist. Assoc.* 77: 568-580.

Haberman, S. J. 1988. A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *J. Amer. Statist. Assoc.* 83: 555-560.

Haberman, S. J. 1995. Computation of maximum likelihood estimates in association models. *J. Amer. Statist. Assoc.* 90:1438-1446.

Hagenaars, J. A. 1998. Categorical causal modeling: Latent class analysis and directed log-linear models with latent variables. *Sociol. Methods Res.* 26: 436-486.

Hald, A. 1998. *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley.

Haldane, J. B. S. 1940. The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small. *Biometrika* 31: 346-355.

Haldane, J. B. S. 1956. The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Human Genet.* 20: 309-311.

Hall, P., and D. M. Titterington. 1987. On smoothing sparse multinomial data. *Austral. J. Statist.* 29: 19-37.

Hamada, M., and C. F. J. Wu. 1990. A critical look at accumulation analysis and related methods. *Technometrics* 32: 119-130.

Hansen, L. P. 1982. Large sample properties of generalized-method of moments estimators. *Econometrica* 50: 1029-1054.

Harkness, W. L., and L. Katz. 1964. Comparison of the power functions for the test of independence in 2 × 2 contingency tables. *Ann. Math. Statist.* 35: 1115-1127.

Harrell F. E., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. 1982. Evaluating the yield of medical tests. *J. Amer. Medic. Assoc.* 247: 2543-2546.

Hartzel, J., I.-M. Liu, and A. Agresti. 2001a. Describing heterogeneous effects in stratified ordinal contingency tables, with application to multi-center clinical trials. *Computat. Statist. Data Anal.* 35: 429-449.

Hartzel, J., A. Agresti, and B. Caffo. 2001b. Multinomial Logit random effects models. *Statistical Modelling* 1: 81-102.

Haslett, S. 1990. Degrees of freedom and parameter estimability in hierarchical models for sparse complete contingency tables. *Computat. Statist. Data Anal.* 9: 179-195.

Hastie, T., and R. Tibshirani. 1987. Non-parametric logistic and proportional odds regression. *Appl. Statist.*

36: 260-276.

Hastie, T., and R. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman & Hall.

Hatzinger, R. 1989. The Rasch model, some extensions and their relation to the class of generalized linear models. *Statistical Modelling: Lecture Notes in Statistics*, Vol. 57. Berlin: Springer-Verlag.

Hauck, W. W. 1979. The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics* 35: 817-819.

Hauck, W. W. 1983. A note on confidence bands for the logistic response curve. *Amer. Statist.* 37: 158-160.

Hauck, W. W., and A. Donner. 1977. Wald's test as applied to hypotheses in Logit analysis. *J. Amer. Statist. Assoc.* 72: 851-853.

Heagerty, P. J. 1999. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* 55: 688-698.

Heagerty, P. J., and S. L. Zeger. 1996. Marginal regression models for clustered ordinal measurements. *J. Amer. Statist. Assoc.* 91: 1024-1036.

Heagerty, P. J., and S. L. Zeger. 2000. Marginalized multilevel models and likelihood inference. *Statist. Sci.* 15: 1-19.

Hedeker, D., and R. D. Gibbons. 1994. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 50: 933-944.

Heinen, T. 1996. *Latent Class and Discrete Latent Trait Models*. Thousand Oaks, CA: Sage Publications.

Heyde, C. C. 1997. *Quasi-likelihood and Its Application*. New York: Springer-Verlag.

Hinde, J. 1982. Compound Poisson regression models. Pp. 109-121 in *GLIM82: Proc. International Conference on Generalised Linear Models*, ed. R. Gilchrist. New York: Springer-Verlag.

Hinde, J., and C. G. B. Demétrio. 1998. Overdispersion: Models and estimation. *Comput. Statist. Data Anal.* 27: 151-170.

Hirji, K. F. 1991. A comparison of exact, mid-*P*, and score tests for matched case-control studies. *Biometrics* 47: 487-496.

Hirji, K. F., C. R. Mehta, and N. R. Patel. 1987. Computing distributions for exact logistic regression. *J. Amer. Statist. Assoc.* 82: 1110-1117.

Hirotsu, C. 1982. Use of cumulative efficient scores for testing ordered alternatives in discrete models. *Biometrika* 69: 567-577.

Hirschfeld, H. O. 1935. A connection between correlation and contingency. *Cambridge Philos. Soc. Proc. (Math. Proc.)* 31: 520-524.

Hodges, J. L., Jr. 1958. Fitting the logistic by maximum likelihood. *Biometrics* 14: 453-461.

Hoem, J. M. 1987. Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review, *Internat. Statist. Rev.* 5: 119-152.

Holford, T. R. 1980. The analysis of rates and of survivorship using log-linear models. *Biometrics* 36: 299-305.

Holt, D., A. J. Scott, and P. D. Ewings. 1980. Chi-squared tests with survey data. *J. Roy. Statist. Soc. Ser. A* 143: 303-320.

Hook, E. B., and R. R. Regal. 1995. Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiol. Rev.* 17: 243-264.

Hosmer, D. W., and S. Lemeshow. 1980. A goodness-of-fit test for multiple logistic regression model. *Commun. Statist. Ser A* 9: 1043-1069.

Hosmer, D. W., and S. Lemeshow. 2000. *Applied Logistic Regression*, 2nd ed. New York: Wiley.

Hosmer, D. W., T. Hosmer, S. le Cessie, and S. Lemeshow. 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Statist. Medic.* 16: 965-980.

Hour, M., O. D. Duncan, and M. E. Sobel. 1987. Association and heterogeneity: Structural models of similarities and differences. *Sociol. Methodol.* 17: 145-184.

Howard, J. V. 1998. The 2 × 2 table: A discussion from a Bayesian viewpoint. *Statist. Sci.* 13: 351-367.

Hsieh, F. Y. 1989. Sample size tables for logistic regression. *Statist. Medic.* 8: 795-802.

Hsieh, F. Y., D. A. Bloch, and M. D. Larsen. 1998. A simple method of sample size calculation for linear and logistic regression. *Statist. Medic.* 17: 1623-1634.

Hwang, J. T. G., and M. T. Wells. 2002. Optimality results for mid *P*-values. To appear.

Hwang, J. T. G., and M.-C. Yang. 2001. An optimality theory for mid *P*-values in 2 × 2 contingency tables. *Statist. Sin.* 11: 807-826.

Imrey, P. B. 1998. Bradley-Terry model. Pp. 437-443 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Imrey, P. B., W. D. Johnson, and G. G. Koch. 1976. An incomplete contingency table approach to paired-comparison experiments. *J. Amer. Statist. Assoc.* 71: 614-623.

Imrey, P. B., G. G. Koch, and M. E. Stokes. 1981. Categorical data analysis: Some reflections on the log linear model and logistic regression. I: Historical and methodological overview. *Internat. Statist. Rev.* 49: 265-283.

Ireland, C. T., and S. Kullback. 1968a. Minimum discrimination information estimation. *Biometrics* 24: 707-713.

Ireland, C. T., and S. Kullback. 1968b. Contingency tables with given marginals. *Biometrika* 55: 179-188.

Ireland, C. T., H. H. Ku, and S. Kullback. 1969. Symmetry and marginal homogeneity of an $r \times r$ contingency table. *J. Amer. Statist. Assoc.* 64: 1323-1341.

Irwin, J. O. 1935. Tests of significance for differences between percentages based on small numbers. *Metron* 12: 83-94.

Jennison, C., and B. W. Turnbull. 2000. *Group Sequential Methods with Applications to Clinical Trials*. London: Chapman & Hall.

Johnson, B. M. 1971. On the admissible estimators for certain fixed sample binomial problems. *Ann. Math. Statist.* 42: 1579-1587.

Johnson, W. 1985. Influence measures for logistic regression: Another point of view. *Biometrika* 72: 59-65.

Johnson, N, L., S. Kotz, and A. W. Kemp. 1992.

Univariate Discrete Distributions, 2nd ed. New York: Wiley.

Jones, B., and M. G. Kenward. 1987. Modelling binary data from a three-period cross-over trial. *Statist. Medic.* 6: 555-564.

Jones, M. P., T. W. O'Gorman, J. H. Lemke, and R. F. Woolson. 1989. A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size considerations. *Biometrics* 45: 171-181.

Jϕrgensen, B. 1983. Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* 70: 19-28.

Jϕrgensen, B. 1987. Exponential dispersion models. *J. Roy. Statist. Soc. Ser. B* 49: 127-162.

Kalbfleisch, J. D., and J. F. Lawless. 1985. The analysis of panel data under a Markov assumption. *J. Amer. Statist. Assoc.* 80: 863-871.

Kastner, C., A. Fieger, and C. Heumann. 1997. MAREG and WinMAREG: A tool for marginal regression models. *Comput. Statist. Data Anal.* 24: 237-241.

Kauermann, G,, and R. J. Carroll, 2001. A note on the efficiency of sandwich covariance matrix estimation. *J. Amer. Statist. Assoc.* 96: 1387-1397.

Kauermann, G., and G. Tutz. 2001. Testing generalized linear and semiparametric models against smooth alternatives. *J. Roy. Statist. Soc. Ser. B* 63: 147-166.

Kelderman, H. 1984. Loglinear Rasch model tests. *Psychometrika* 49: 223-245.

Kempthorne, O. 1979. In dispraise of the exact test: Reactions. *J. Statist. Plann. Inference* 3: 199-213.

Kendall, M. G. 1945. The treatment of ties in rank problems. *Biometrika* 33: 239-251.

Kendall, M., and A. Stuart. 1979. *The Advanced Theory of Statistics*, Vol. 2; *Inference and Relationship*, 4th ed. New York: Macmillan.

Kenward, M. G., and B. Jones. 1991. The analysis of categorical data from cross-over trials using a latent variable model. *Statist. Medic.* 10: 1607-1619.

Kenward, M. G., and B. Jones. 1994. The analysis of binary and categorical data from crossover trials. *Statist. Methods Medic. Res.* 3: 325-344.

Kenward, M. G., E. Lesaffre, and G. Molenberghs. 1994. An application of maximum likelihood and estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* 50: 945-953.

Khamis, H. J. 1983. Log-linear model analysis of the semi-symmetric intraclass contingency table. *Commun. Statist. Ser. A* 12: 2723-2752.

Kim, D., and A. Agresti. 1995. Improved exact inference about conditional association in three-way contingency tables. *J. Amer. Statist. Assoc.* 90: 632-639.

Kim, D., and A. Agresti. 1997. Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables. *Comput. Statist. Data Anal.* 24: 89-104.

King, G. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton University Press.

Knuiman, M. W., and T. P. Speed. 1988. Incorporating prior information into the analysis of contingency tables. *Biometrics* 44: 1061-1071.

Koch, G. G., and V. P. Bhapkar. 1982. Chi-square tests. Pp. 442-457 in *Encyclopedia of Statistical Sciences*, Vol. 1. New York: Wiley.

Koch, G. G., J. R. Landis, J. L. Freeman, D. H. Freeman, and R. G. Lehnen. 1977. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* 33: 133-158.

Koch, G. G., I. A. Amara, G. W. Davis, and D. B. Gillings. 1982. A review of some statistical methods for covariance analysis of categorical data. *Biometrics* 38: 563-595.

Koch, G. G., P. B. Imrey, J. M. Singer, S. S. Atkinson, and M. E. Stokes. 1985. *Lecture Notes for Analysis of Categorical Data*. Montreal: Les Presses de L'Université de Montréal.

Koehler, K. 1986. Goodness-of-fit tests for log-linear models in sparse contingency tables. *J. Amer. Statist. Assoc.* 81: 483-493.

Koehler, K. 1998. Chi-square tests. Pp. 608-622 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Koehler, K., and K. Larntz. 1980. An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Amer. Statist. Assoc.* 75: 336-344.

Koehler, K., and J. Wilson. 1986. Chi-square tests for comparing vectors of proportions for several cluster samples. *Commun. Statist. Ser. A* 15: 2977-2990.

Koopman, P. A. R. 1984. Confidence limits for the ratio of two binomial proportions. *Biometrics* 40: 513-517.

Kraemer, H. C. 1979. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* 44: 461-472.

Kreiner, S. 1987. Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies. *Scand. J. Statist.* 14: 97-112.

Kreiner, S. 1998. Interaction models. Pp. 2063-2068 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Kruskal, W. H. 1958. Ordinal measures of association. *J. Amer. Statist. Assoc.* 53: 814-861.

Ku, H. H., R. N. Varner, and S. Kullback. 1971. Analysis of multidimensional contingency tables. *J. Amer. Statist. Assoc.* 66: 55-64.

Kuha, J., and C. Skinner. 1997. Categorical data analysis and misclassification. Pp. 633-670 in *Survey Measurement and Process Quality*, ed. L. Lyberg et al. New York: Wiley.

Kuha, J., C. Skinner, and J. Palmgren. 1998. Misclassification error. Pp. 2615-2621 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Kullback, S. 1959. *Information Theory and Statistics*. New York: Wiley.

Kullback, S., M. Kupperman, and H. H. Ku. 1962. Tests for contingency tables and Markov chains. *Technometrics* 4: 573-608.

Kupper, L. L., and J. K. Haseman. 1978. The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* 34: 69-76.

Kupper, L. L., C. Portier, M. D. Hogan, and E. Yamamoto. 1986. The impact of litter effects on dose-response modeling in teratology. *Biometrics* 42: 85-98.

Läärä, E., and J. N. S. Matthews. 1985. The equivalence of two models for ordinal data. *Biometrika* 72: 206-207.

Lachin, J. M. 1977. Sample-size determinations for $r \times c$ comparative trials. *Biometrics* 33: 315-324.

Laird, N. M. 1978. Empirical Bayes methods for two-way contingency tables. *Biometrika* 65: 581-590.

Laird, N. M. 1998. EM algorithm. Pp. 1300-1313 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Laird, N. M., and D. Olivier. 1981. Covariance analysis of censored survival data using log-linear analysis techniques. *J. Amer. Statist. Assoc.* 76: 231-240.

Lancaster, H. O. 1949. The derivation and partition of χ^2 in certain discrete distributions. *Biometrika* 36: 117-129.

Lancaster, H. O. 1951. Complex contingenly tables treated by partition of χ^2. *J. Roy. Statist. Soc. Ser. B* 13: 242-249.

Lancaster, H. O. 1961. Significance tests in discrete distributions. *J. Amer. Statist. Assoc.* 56: 223-234.

Lancaster, H. O. 1969. *The Chi-Squared Distribution*. New York: Wiley.

Lancaster, H. O., and M, A. Hamdan. 1964. Estimation of the correlation coefficient in contingency tables with possible nonmetrical characters. *Psychometrika* 29: 383-391.

Landis, J. R., and G. G. Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 33: 363-374.

Landis, J. R., E. R. Heyman, and G. G. Koch. 1978. Average partial association in three-way contingency tables: A review and discussion of alternative tests. *Internat. Statist. Rev.* 46: 237-254.

Landis, J. R., T. J. Sharp, S. J. Kuritz, and G. G. Koch. 1998. Mantel-Haenszel methods. Pp. 2378-2691 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Landwehr, J. M., D. Pregibon, and A. C. Shoemaker. 1984. Graphical methods for assessing logistic regression models. *J. Amer. Statist. Assoc.* 79: 61-71.

Lang, J. B. 1992. Obtaining the observed information matrix for the Poisson log linear model with incomplete data. *Biometrika* 79: 405-407.

Lang, J. 13. 1996a. Maximum likelihood methods for a generalized class of log-linear models. *Ann. Statist.* 24: 726-752.

Lang, J. B. 1996b. On the partitioning of goodness-of-fit statistics for multivariate categorical response models. *J. Amer. Statist. Assoc.* 91: 1017-1023.

Lang, J. B. 1996c. On the comparison of multinomial and Poisson log-linear models. *J. Roy. Statist. Soc. Ser. B* 58: 253-266.

Lang, J. B., and A. Agresti. 1994. Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* 89: 625-632.

Lang, J. B., J. W. McDonald, and P. W. F. Smith. 1999. Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach. *J. Amer. Statist. Assoc.* 94: 1161-1171.

Laplace, P. S. 1812. *Théorie Analytique des Probabilités*. Paris: Courcier.

Larntz, K. 1978. Small-sample comparison of exact levels for chi-squared goodness-of-fit statistics. *J. Amer. Statist. Assoc.* 73: 253-263.

Larsen, K., J. H. Petersen, E. Budtz-Jørgensen, and L. Endahl. 2000. Interpreting parameters in the logistic regression model with random effects. *Biometrics* 56: 909-914.

Larson, M. G. 1984. Covariate analysis of competing-risks data with log-linear models. *Biometrics* 40: 459-469.

Lauritzen, S. L. 1996. *Graphical Models*. New York: Oxford University Press.

Lauritzen, S, L., and N. Wermuth, 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist*, 17: 31-57.

LaVange, L. M., G. G. Koch, and T. A. Schwartz. 2001. Applying sample survey methods to clinical trials data. *Statist. Medic.* 20: 2609-2623.

Lawal, H. B. 1984. Comparisons of the X^2, Y^2, Freeman-Tukey and Williams improved G^2 test statistics in small samples of one-way multinomials. *Biornetrika* 71: 415-418.

Lawless, J. F. 1987. Negative binomial and mixed Poisson regression, *Canad. J. Statist.* 15: 209-225.

Lazarsfeld, P. F., and N. W. Henry. 1968. *Latent Structure Analysis*. Boston: Houghton Mifflin.

Lee, S. K. 1977. On the asymptotic variances of û terms in loglinear models of multidimensional contingency tables. *J. Amer. Statist. Assoc.* 72: 412-419.

Lee, Y., and J. A. Nelder. 1996. Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser B* 58: 619-678.

Lefkopoulou, M., D. Moore, and L. Ryan. 1989. The analysis of multiple correlated binary outcomes: Application to rodent teratology experiments. *J. Amer. Statist. Assoc.* 84: 810-815.

Lehmann, E. L. 1966. Some concepts of dependence. *Ann. Math. Statist.* 37: 1137-1153.

Lehmann, E. L. 1986. *Testing Statistical Hypotheses*, 2nd ed. New York: Wiley.

Leonard, T. 1975. Bayesian estimation methods for two-way contingency tables. *J. Roy. Statist. Soc. Ser. B* 37: 23-37.

Leonard, T. and J. S. J. Hsu. 1994. The Bayesian analysis of categorical data: A selective review. Pp. 283-310 in *Aspects of Uncertainty: A Tribute to D. V. Lindley*. P. R. Freeman and A. F. M. Smith, eds. New York: Wiley.

Lesaffre, E., and A. Albert. 1989. Multiple-group logistic regression diagnostics. *Appl. Statist.* 38: 425-440.

Lesaffre, E., and G. Molenberghs. 1991. Multivariate probit analysis: A neglected procedure in medical statistics. *Statist. Medic.* 10: 1391-1403.

Lesaffre, E., and B. Spiessens. 2001. On the effect of quadrature points in a logistic random-effects model: An example. *Appl. Statist.* 50: 325-335.

Lewis, T., I. W. Saunders, and M. Westcott. 1984. The moments of the Pearson chi-squared statistic and the minimum expected value in two-way tables. *Biometrika*

71: 515-522.

Liang, K. Y. 1984. The asymptotic efficiency of conditional likelihood methods. *Biometrika* 71:305-313.

Liang, K. Y., and J. Hanfelt. 1994. On the use of the quasi-likelihood method in teratological experiments. *Biometrics* 50: 872-880.

Liang, K. Y., and P. McCullagh. 1993. Case studies in binary dispersion. *Biometrics* 49: 623-630.

Liang, K. Y., and S. G. Self. 1985. Tests for homogeneity of odds ratios when the data are sparse. *Biometrika* 72: 353-358.

Liang, K. Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13-22.

Liang, K. Y., and S. L. Zeger. 1988. On the use of concordant pairs in matched case-control studies. *Biometrics* 44: 1145-1156.

Liang, K. Y, and S. L. Zeger. 1995. Inference based on estimating functions in the presence of nuisance parameters. *Statist. Sci.* 10 158-173.

Liang, K. Y., S. L. Zeger, and B. Qaqish. 1992. Multivariate regression analyses for categorical data. *J. Roy. Statist. Soc. Ser. B* 54: 3-24.

Lin, X. 1997. Variance component testing in generalized linear models with random effects. *Biometrika* 84: 309-326.

Lindley, D. V. 1964. The Bayesian analysis of contingency tables. *Ann. Math. Statist.* 35: 1622-1643.

Lindsay, B., C. Clogg, and J. Grego. 1991. Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Amer. Statist. Assoc.* 86: 96-107.

Lindsey, J. K. 1999. *Models for Repeated Measurements*, 2nd ed. Oxford: Oxford University Press.

Lindsey, J. K., and P. M. E. Altham. 1998. Analysis of the human sex ratio by using overdispersion models. *Appl. Statist.* 47: 149-157.

Lindsey, J. K., and G. Mersch. 1992. Fitting and comparing probability distributions with log linear models. *Comput. Statist. Data Anal.* 13: 373-384.

Lipsitz, S. 1992. Methods for estimating the parameters of a linear model for ordered categorical data. *Biometrics* 48: 271-281.

Lipsitz, S. R., and G. Fitzmaurice. 1996. The score test for independence in $R \times C$ contingency tables with missing data. *Biometrics* 52: 751-762.

Lipsitz, S., N. Laird, and D. Harrington. 1990. Finding the design matrix for the marginal homogeneity model. *Biometrika* 77: 353-358.

Lipsitz, S., N. Laird, and D. Harrington. 1991. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* 78: 153-160.

Lipsitz, S. R., K. Kim, and L. Zhao. 1994. Analysis of repeated categorical data using generalized estimating equations. *Statist. Medic.* 13: 1149-1163.

Little, R. J. 1989. Testing the equality of two independent binomial proportions. *Amer. Statist.* 43: 283-288.

Little, R. J. 1998. Missing data. Pp. 2622-2635 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Little, R. J., and D. B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R. J. A., and M.-M. Wu. 1991. Models for contingency tables with known margins when target and sampled populations differ. *J. Amer. Statist. Assoc.* 86: 87-95.

Liu, Q., and D. A. Pierce. 1993. Heterogeneity in Mantel-Haenszel-type models. *Biometrika* 80: 543-556.

Liu, Q., and D. A. Pierce. 1994. A note on Gauss-Hermite quadrature. *Biometrika* 81: 624-629.

Lloyd, C. J. 1988a. Some issues arising from the analysis of 2 × 2 contingency tables. *Austral. J. Statist.* 30: 35-46.

Lloyd, C. J. 1988b. Doubling the one-sided *P*-value in testing independence in 2 × 2 tables against a two-sided alternative. *Statist. Medic.* 7: 1297-1306.

Lloyd, C. J. 1999. *Statistical Analysis of Categorical Data*. New York: Wiley.

Longford, N. T. 1993. *Random Coefficient Models*. New York: Oxford University Press.

Loughin, T. M., and P. N. Scherer. 1998. Testing for association in contingency tables with multiple column responses. *Biometrics* 54: 630-637.

Louis, T. A. 1982. Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 44: 226-233.

Luce, R. D. 1959. *Individual Choice Behavior*. New York: Wiley.

Madansky, A. 1963. Tests of homogeneity for correlated samples. *J. Amer. Statist. Assoc.* 58: 97-119.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Magnus, J. R., and H. Neudecker. 1988. *Matfix Differential Calculus with Applications in Statistics and Econometrics*. New York: Wiley.

Mantel, N. 1963. Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *J. Amer. Statist. Assoc.* 58: 690-700.

Mantel, N. 1966. Models for complex contingency tables and polychotomous dosage response curves. *Biometrics* 22: 83-95.

Mantel, N. 1973. Synthetic retrospective studies and related tppics. *Biometrics* 29: 479-486.

Mantel, N. 1985. Maximum likelihood vs. minimum chi-square. *Biometrics* 41: 777-781.

Mantel, N. 1987a. Understanding Wald ' s test for exponential families. *Amer. Statist.* 41: 147-148.

Mantel, N. 1987b. Exact tests for 2 × 2 contingency tables (Letter). *Amer. Statist.* 41: 159.

Mantel, N., and D. P. Byar. 1978. Marginal homogeneity, symmetry and independence. *Commun. Statist. Ser. A* 7: 953-976.

Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* 22: 719-748.

Martin Andrés, A., and Silva Mato, A. 1994. Choosing the optimal unconditional test for comparing two independent proportions. *Comput. Statist. Data Anal.* 17: 555-574.

Matthews, J. N. S., and K. P. Morris. 1995. An application of Bradley-Terry-type models to the measurement of pain. *Appl. Statist.* 44: 243-255.

McCullagh, P. 1978. A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika* 65: 413-418.

McCullagh, P. 1980. Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B* 42: 109-142.

McCullagh, P. 1982. Some applications of quasisymmetry. *Biometrika* 69: 303-308.

McCullagh, P. 1983. Quasi-likelihood functions. *Ann. Statist.* 11: 59-67.

McCullagh, P. 1986. The conditional distribution of goodness-of-fit statistics for discrete data. *J. Amer. Statist. Assoc.* 81: 104-107.

McCullagh, P., and J. A. Nelder. 1983; 2nd ed., 1989. *Generalized Linear Models*. London: Chapman & Hall.

McCulloch, C. E. 1994. Maximum likelihood variance components estimation for binary data. *J. Amer. Statist. Assoc.* 89: 330-335.

McCulloch, C. E. 1997. Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* 92: 162-170.

McCulloch, C. E. 2000. Generalized linear models. *J. Amer. Statist. Assoc.* 95: 1320-1324.

McCulloch, C. E., and S. Searle. 2001. *Generalized, Linear, and Mixed Models*. New York: Wiley.

McFadden, D. 1974. Conditional Logit analysis of qualitative choice behavior. Pp. 105-142 in *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

McFadden, D. 1982. Qualitative response models. Pp. 1-37 in *Advances in Econometrics*, ed. W. Hildebrand. Cambridge: Cambridge University Press.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12: 153-157.

Mee, R. W. 1984. Confidence bounds for the difference between two probabilities (letter). *Biometrics* 40: 1175-1176.

Meeden, G., C. Geyer, J. Lang, and E. Funo. 1998. The admissibility of the maximum likelihood estimator for decomposable log-linear interaction models for contingency tables. *Commun. Statist. Ser. A* 27: 473-493.

Mehta, C. R. 1994. The exact analysis of contingency tables in medical research. *Statist. Methods Medic. Res.* 3: 135-156.

Mehta, C. R., and N. R. Patel. 1983. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Amer. Statist. Assoc.* 78: 427-434.

Mehta, C. R., and N. R. Patel. 1995. Exact logistic regression: Theory and examples. *Statist. Medic.* 14: 2143-2160.

Mehta, C. R., and S. J. Walsh. 1992. Comparison of exact, mid-*P*, and Mantel-Haenszel confidence intervals for the common odds ratio across several 2 ×2 contingency tables. *Amer. Statist.* 46: 146-150.

Mehta, C. R., N. R. Patel, and R. Gray. 1985. Computing an exact confidence interval for the common odds ratio in several 2 by 2 contingency tables. *J. Amer. Statist. Assoc.* 80: 969-973.

Mehta, C. R., N. R. Patel, and P. Senchaudhuri. 1988. Importance sampling for estimating exact probabilities in permutational inference. *J. Amer. Statist. Assoc.* 83: 999-1005.

Mehta, C. R., N. R. Patel, and P. Senchaudhuri. 2000. Efficient Monte Carlo methods for conditional logistic regression. *J. Amer. Statist. Assoc.* 95: 99-108.

Michailidis, G., and J. de Leeuw. 1998. The Gifi system of descriptive multivariate analysis. *Statist. Sci.* 13: 307-336.

Miettinen, O. S. 1969. Individual matching with multiple controls in the case of all-or-none responses. *Biometrics* 25: 339-355.

Miettinen, O. S., and M. Nurminen. 1985. Comparative analysis of two rates. *Statist. Medic.* 4: 213-226.

Miller, M. E., C. S. Davis, and J. R. Landis. 1993. The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics* 49: 1033-1044.

Minkin, S. 1987. On optimal design for binary data. *J. Amer. Statist. Assoc.* 82: 1098-1103.

Mirkin, B. 2001. Eleven ways to look at the chi-squared coefficient for contingency tables. *Amer. Statist.* 55: 111-120.

Mitra, S. K. 1958. On the limiting power function of the frequency chi-square test. *Ann. Statist.* 29: 1221-1233.

Molenberghs, G., and E. Goetghebeur. 1997. Simple fitting algorithms for incomplete categorical data. *J. Roy. Statist. Soc. Ser. B* 59: 401-414.

Molenberghs, G., and E. Lesaffre. 1994. Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *J. Amer, Statist. Assoc.* 89: 633-644.

Molenberghs, G., M. G. Kenward, and E. Lesaffre. 1997. The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika* 84: 33-44.

Moore, D. F. 1986a. Asymptotic properties of moment estimates for overdispersed counts and proportions. *Biometrika* 35: 583-588.

Moore, D. S. 1986b. Tests of chi-squared type. Pp. 63-95 in *Goodness-of-Fit Techniques*, ed. R. D' Agostino and M. A. Stephens. New York: Marcel Dekker.

Moore, D. F., and A. Tsiatis. 1991. Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics* 47: 383-401.

Morgan, B. J. T. 1992. *Analysis of Quantal Response Data*. London: Chapman & Hall.

Morgan, W. M., and B. A. Blumenstein. 1991. Exact conditional tests for hierarchical models in multidimensional contingency tables. *Appl. Statist.* 40: 435-442.

Mosimann, J. E. 1962. On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions. *Biometrika* 49: 65-82.

Mosteller, F. 1951. Remarks on the method of paired comparisons I: The least-squares solution assuming equal standard deviations and equal correlations. *Psychometrika* 16: 3-9.

Mosteller, F. 1952. Some statistical problems in measuring the subjective response to drugs. *Biometrics* 8: 220-226.

Mosteller, F. 1968. Association and estimation in contingency tables. *J. Amer. Statist. Assoc.* 63: 1-28.

Naif, V. N. 1987. Chi-squared-type tests for ordered alternatives in contingency tables. *J. Amer. Statist. Assoc.* 82: 283-291.

Natarajan, R., and C. McCulloch. 1995. A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* 82: 639-643.

Natarajan, R., and C. McCulloch. 1998. Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *J. Comput. Graph. Statist.* 7: 267-277.

Nelder, J., and D. Pregibon. 1987. An extended quasi-likelihood function. *Biometrika* 74: 221-232,

Nelder, J., and R. W. M. Wedderburn. 1972. Generalized linear models. *J. Roy. Statist. Soc. Ser. A* 135: 370-384.

Nerlove, M., and S. J. Press. 1973. Univariate and multivariate log-linear and logistic models. Technical Report R-1306-EDA/NIH, Rand Corporation, Santa Monica, CA.

Neuhaus, J. M. 1992. Statistical methods for longitudinal and clustered designs with binary responses. *Statist. Methods Medic. Res.* 1: 249-273.

Neuhaus, J. M., and N. P. Jewell. 1990a. Some comments on Rosner's multiple logistic model for clustered data. *Biometrics* 46: 523-534.

Neuhaus, J. M., and N. P. Jewell. 1990b. The effect of retrospective sampling on binary regression models for clustered data. *Biometrics* 46: 977-990.

Neuhaus, J. M., and M. L. Lesperance. 1996. Estimation efficiency in a binary mixed-effects model setting. *Biometrika* 83: 441-446.

Neuhaus, J. M., J. D. Kalbfleisch, and W. W. Hauck. 1991. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Internat. Statist. Rev.* 59: 25-35.

Neuhaus, J. M., W. W. Hauck, and J. D. Kalbfleisch. 1992. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* 79: 755-762.

Neuhaus, J. M., J. D. Kalbfleisch, and W. W. Hauck. 1994. Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *Canad. J. Statist.* 22: 139-148.

Newcombe, R. 1998a. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statist. Medic.* 17: 857-872.

Newcombe, R. 1998b. Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statist. Medic.* 17: 873-890.

Newcombe, R. 2001. Logit confidence intervals and the inverse sinh transformation. *Amer. Statist.* 55: 200-202.

Neyman, J. 1935. On the problem of confidence limits. *Ann. Math. Statist.* 6: 111-116.

Neyman, J. 1949. Contributions to the theory of the χ^2 test. Pp. 239-273 in *Proc. First Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley, CA: University of California Press.

Nurminen, M. 1986. Confidence intervals for the ratio and difference of two binomial proportions. *Biometrics* 42: 675-676.

O'Brien, P. C. 1988. Comparing two samples: Extensions of the t, rank-sum, and log-rank tests. *J. Amer. Statist. Assoc.* 83: 52-61.

O'Brien, R. G. 1986. Using the SAS system to perform power analyses for log-linear models. Pp. 778-784 in *Proc. 11th Annual SAS Users Group Conference*. Cary, NC: SAS Institute.

Ochi, Y., and R. Prentice. 1984. Likelihood inference in a correlated probit regression model. *Biometrika* 71: 531-543.

O'Gorman, T. W., and R. F. Woolson. 1988. Analysis of ordered categorical data using the SAS system. Pp. 957-963 in *Proc. 13th Annual SAS Users Group Conference*. Cary, NC: SAS Institute.

Paik, M. 1985. A graphic representation of a three-way contingency table: Simpson's paradox and correlation. *Amer. Statist.* 39: 53-54.

Palmgren, J. 1981. The Fisher information matrix for log-linear models arguing conditionally in the observed explanatory variables. *Biometrika* 68: 563-566.

Palmgren, J., and A. Ekholm. 1987. Exponential family non-linear models for categorical data with errors of observation. *Appl. Stochastic Models Data Anal.* 3: 111-124.

Park, T., and M. B. Brown. 1994. Models for categorical data with nonignorable nonresponse. *J. Amer. Statist. Assoc.* 89: 44-52.

Parr, W. C., and H. D. Tolley. 1982. Jackknifing in categorical data analysis. *Austral. J. Statist.* 24: 67-79.

Parzen, E. 1997. Concrete statistics. Pp. 309-332 in *Statistics of Quality*. New York: Marcel Dekker.

Patefield, W. M. 1982. Exact tests for trends in ordered contingency tables. *Appl. Statist. Ser B* 31: 32-43.

Patnaik, P. B. 1949. The non-central χ^2 and F-distributions and their applications. *Biometrika* 36: 202-232.

Paul, S. R., K. Y. Liang, and S. G. Self. 1989. On testing departure from the binomial and multinomial assumptions. *Biometrics* 45: 231-236.

Pearson, E. S. 1947. The choice of a statistical test illustrated on the interpretation of data classified in 2 × 2 tables. *Biometrika* 34: 139-167.

Pearson, K. 1900. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag. Ser.* 5 50: 157-175. (Reprinted in *Karl Pearson's Early Statistical Papers*, ed. E. S. Pearson. Cambridge: Cambridge University Press, 1948.)

Pearson, K. 1904. Mathematical contributions to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation. *Draper's Co. Research Memoirs, Biometric Series*, no. 1. (Reprinted *in Karl Pearson's Early Papers*, ed. E. S.

Pearson, Cambridge: Cambridge University Press, 1948.)

Pearson, K. 1913. On the probable error of a correlation coefficient as found from a fourfold table. *Biometrika* 9: 22-27.

Pearson, K. 1917. On the general theory of multiple contingency with special reference to partial contingency. *Biometrika* 11: 145-158.

Pearson, K. 1922. On the χ^2 test of goodness of fit. *Biometrika* 14: 186-191.

Pearson, K., and D. Heron. 1913. On theories of association. *Biometrika* 9: 159-315.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. 1996. A simulation study of the number of events per variable in logistic regression ȁnalysis. *J. Clin. Epidemiol.* 49: 1373-1379.

Pendergast, J. F., S. J. Gange, M. A. Newton, M. J. Lindstrom, M. Palta, and M. R. Fisher. 1996. A survey of methods for analyzing clustered binary response data. *Internat. Statist. Rev.* 64: 89-118.

Pepe, M. S. 2000. Receiver operating characteristic methodology. *J. Amer. Statist. Assoc.* 95: 308-311.

Peterson, B., and F. E. Harrell, Jr. 1990. Partial proportional odds models for ordinal response variables. *Appl. Statist.* 39: 205-217.

Pierce, D. A., and D. Peters. 1992. Practical use of higher order asymptotics for multiparameter exponential families. *J. Roy. Statist. Soc. Ser. B* 54: 701-725.

Pierce, D. A., and D. Peters. 1999. Improving on exact tests by approximate conditioning. *Biometrika* 86: 265-277.

Pierce, D. A., and B. R. Sands. 1975. Extra-Bernoulli variation in regression of binary data. Technical Report 46, Statistics Deptartment, Oregon State University, Cornwallis, OR.

Pierce, D. A., and D. W. Schafer. 1986. Residuals in generalized linear models. *J. Amer. Statist. Assoc.* 81: 977-983.

Plackett, R. L. 1962. A note on interactions in contingency tables. *J. Roy. Statist. Soc. Ser. B* 24: 162-166.

Plackett, R. L. 1964. The continuity correction in 2 × 2 tables. *Biometrika* 51: 327-337.

Plackett, R. L. 1983. Karl Pearson and the chi-squared test. *Internat. Statist. Rev.* 51: 59-72.

Podgor, M. J., J. L. Gastwirth, and C. R. Mehta. 1996. Efficiency robust tests of independence in contingency tables with ordered classifications. *Statist. Medic.* 15: 2095-2105.

Poisson, S.-D. 1837. *Recherches sur la probabilité des jugements en matieère criminelle et en matière civile, précédées des règles générales du calcul des probabilités.* Paris: Bachelier.

Pratt, J. W. 1981. Concavity of the log likelihood. *J. Amer. Statist. Assoc.* 76: 103-106.

Pregibon, D. 1980. Goodness of link tests for generalized linear models. *Appl. Statist.* 29: 15-24.

Pregibon, D. 1981. Logistic regression diagnostics. *Ann. Statist.* 9: 705-724.

Pregibon, D. 1982. Score tests in GLIM with application. Pp. 87-97 in *Lecture Notes in Statistics*, 14: *GLIM 82, Proc. International Conference on Generalised Linear Models*, ed. R. Gilchrist. New York: Springer-Verlag.

Prentice, R. 1976a. Use of the logistic model in retrospective studies. *Biometrics* 32: 599-606.

Prentice, R. 1976b. Generalization of the probit and Logit methods for dose response curves. *Biometrics* 32: 761-768.

Prentice, R. 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Amer. Statist. Assoc.* 81: 321-327.

Prentice, R., and N. Breslow. 1978. Retrospective studies and failure time models. *Biometrika* 65: 153-158.

Prentice, R., and L. A. Gloeckler. 1978. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 34: 57-67.

Prentice, R., and R. Pyke. 1979. Logistic disease incidence models and case-control studies. *Biometrika* 66: 403-412.

Prentice, R., and L. P. Zhao. 1991. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* 47: 825-839.

Press, S. J., and S. Wilson. 1978. Choosing between logistic regression and discriminant analysis. *J. Amer. Statist. Assoc.* 73: 699-705.

Qu, A., B. G. Lindsay, and B. Li. 2000. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 87: 823-836.

Quine, M. P., and E. Seneta. 1987. Bortkiewicz's data and the law of small numbers, *Internat. Statist. Rev.* 5: 173-181.

Rabe-Hesketh, S., and A. Skrondal. 2001. Parameterisation of multivariate random effects models for categorical data. *Biometrics* 57: -.

Raftery, A. E. 1986. Choosing models for cross-classification. *Amer. Sociol. Rev.* 51: 145-146.

Rao, C. R. 1957. Maximum likelihood estimation for the multinomial distribution. *Sankhya* 18: 139-148.

Rao, C. R. 1963. Criteria of estimation in large samples. *Sankhya* 25: 189-206.

Rao, C. R. 1973. *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley.

Rao, C. R. 1982. Diversity: Its measurement, decomposition, apportionment, and analysis. *Sankhya Ser. A* 44: 1-22.

Rao, J. N. K., and A. J. Scott. 1987. On simple adjustments to chi-square tests with sample survey data. *Ann. Statist.* 15: 385-397.

Rao, J. N. K., and D. R. Thomas. 1988. The analysis of cross-classified categorical data from complex sample surveys. *Sociol. Methodol.* 18: 213-270.

Rasch, G. 1961. On general laws and the meaning of measurement Jn psychology. Pp. 321-333 in *Proc. 4th Berkeley Symposium on Mathematics, Statistics, and Probability*, Vol. 4, ed. J. Neyman. Berkeley, CA: University of California Press.

Rayner, J. C. W., and D. J. Best. 2001. *A Contingency Table Approach to Nonparametric Testing.* London: Chapman & Hall.

Read, T. R. C., and N. A. C. Cressie. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data.* New

York: Springer-Verlag.

Rice, W. R. 1988. A new probability model for determining exact P-values for 2×2 contingency tables when comparing binomial proportions. *Biometrics* 44: 1-22.

Ritov, Y., and Z. Gilula. 1991. The order-restricted RC model for ordered contingency tables: Estimation and testing for fit. *Ann. Statist.* 19: 2090-2101.

Robins, J., N. Breslow, and S. Greenland. 1986. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 42: 311-323.

Robins, J., A. Rotnitzky, and L. P. Zhao. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* 90: 106-121.

Röhmel, J., and U. Mansmann. 1999. Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical J.* 41: 149-170.

Rosenbaum, P. R., and D. R. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41-55.

Rosner, B. 1984. Multivariate methods in ophthalmology with application to other paired-data situations. *Biometrics* 40: 1025-1035.

Rosner, B. 1989. Multivariate methods for clustered binary data with more than one level of nesting. *J. Amer. Statist. Assoc.* 84: 373-380.

Rotnitzky, A., and N. P. Jewell. 1990. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 77: 485-497.

Routledge, R. D. 1992. Resolving the conflict over Fisher's exact test. *Canad. J. Statist.* 20: 201-209.

Routledge, R. D. 1994. Practicing safe statistics with the mid-P*. *Canad. J. Statist.* 22: 103-110.

Roy, S. N., and M. A. Kastenbaum. 1956. On the hypothesis of no "interaction" in a multiway contingency table. *Ann. Math. Statist.* 27: 749-757.

Roy, S. N., and S. K. Mitra. 1956. An introduction to some nonparametric generalizations of analysis of variance and multivariate analysis. *Biometrika* 43: 361-376.

Rudas, T., C. C. Clogg, and B. G. Lindsay. 1994. A new index of fit based on mixture methods for the analysis of contingency tables. *J. Roy. Statist. Soc.* 56: 623-639.

Ryan, L. 1992. Quantitative risk assessment for developmental toxicity. *Biometrics* 48: 163-174.

Ryan, L. 1995. Comment on article by Liang and Zeger. *Statist. Sci.* 10: 189-193.

Samuels, M. L. 1993. Simpson's paradox and related phenomena. *J. Amer. Statist. Assoc.* 88: 81-88.

Santner, T. J., and M. K. Snell. 1980. Small-sample confidence intervals for P_1-P_2 and P_1/P_2 in 2×2 contingency tables. *J. Amer. Statist. Assoc.* 75: 386-394.

Santner, T. J., and S. Yamagami. 1993. Invariant small sample confidence intervals for the difference of two success probabilities. *Commun. Statist. Ser. B* 22: 33-59.

Schafer, J. L. 1997. *Analysis of Ineomplete Multiuariate Data*. London: Chapman & Hall.

Schluchter, M. D., and K. L. Jackson. 1989. Log-linear analysis of censored survival data with partially observed covariates. *J. Amer. Statist. Assoc.* 84: 42-52.

Scott, A., and C. Wild. 2001. Case-control studies with complex sampling. *Appl. Statist.* 50: 389-401.

Seeber, G. 1998. Poisson regression. Pp. 3404-3412 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Sekar, C. C., and W. E. Deming. 1949. On a method of estimating birth and death rates and the extent of registration. *J. Amer. Statist. Assoc.* 44: 101-115.

Self, S. G., and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* 82: 605-610.

Sen, P. K., and J. M. Singer. 1993. *Large Sample Methods in Statistics: An Introduction with Applications*. London: Chapman & Hall.

Shapiro, S. H. 1982. Collapsing contingency tables: A geometric approach. *Amer. Statist.* 36: 43-46.

Shuster, J., and D. Downing. 1976. Two-way contingency tables for complex sampling schemes. *Biometrika* 63: 271-276.

Silvapulle, M. J. 1981. On the existence of maximum likelihood estimators for the binomial response models. *J. Roy. Statist. Soc. Ser. B* 43: 310-313.

Simon, G. 1973. Additivity of information in exponential family probability laws. *J. Amer. Statist. Assoc.* 68: 478-482.

Simon, G. 1974. Alternative analyses for the singly-ordered contingency table. *J. Amer. Statist. Assoc.* 69: 971-976.

Simon, G. 1978. Efficacies of measures of association for ordinal contingency tables. *J. Amer. Statist. Assoc.* 73: 545-551.

Simonoff, J. 1983. A penalty function approach to smoothing large sparse contingency tables. *Ann. Statist.* 11: 208-218.

Simonoff, J. 1986. Jackknifing and bootstrapping goodness-of-fit statistics in sparse multinomials. *J. Amer. Statist. Assoc.* 81: 1005-1111.

Simonoff, J. S. 1996. *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Simonoff, J. S. 1998. Three sides of smoothing: Categorical data smoothing, nonparametric regression, and density estimation. *Internat. Statist. Rev.* 66: 137-156.

Simpson, E. H. 1949. The measurement of diversity. *Nature* 163: 699.

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* 13: 238-241.

Skellam, J. G. 1948. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *J. Roy. Statist. Soc. Ser. B* 10: 257-261.

Skene, A. M., and J. C. Wakefield. 1990. Hierarchical models for multicentre binary response studies. *Statist. Medic.* 9: 919-929.

Slaton, T. L., W. W. Piegorsch, and S. D. Durham. 2000. Estimation and testing with overdispersed proportions using the beta-logistic regression model of Heckman and Willis. *Biometrics* 56: 125-133.

Small, K. A. 1987. A discrete choice model for ordered alternatives. *Econometrica* 55: 409-424.

Smith, K. W. 1976. Table standardization and table shrinking: Aids in the traditional analysis of contingency tables. *Social Forces* 54: 669-693.

Smith, P. W. F., J. J. Forster, and J. W. McDonald. 1996. Monte Carlo exact tests for square contingency tables. *J. Roy. Statist. Soc. Ser. A* 159: 309-321.

Snell, E. J. 1964. A scaling procedure for ordered categorical data. *Biometrics* 20: 592-607.

Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. *Amer. Sociol. Rev.* 27: 799-811.

Speed, T. 1998. Iterative proportional fitting. Pp. 2116-2119 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Spiegelhalter, D. J., and A. F. M. Smith. 1982. Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. Ser. B* 44: 377-387.

Spitzer, R. L., J. Cohen, J. L. Fleiss, and J. Endicott. 1967. Quantification of agreement in psychiatric diagnosis. *Arch. Gen. Psychiatry* 17: 83-87.

Sprott, D. A. 2000. *Statistical Inference in Science*. New Yorki Springer-Verlag.

Stern, S. 1997. Simulation-based estimation. *J. Econ. Literature* 35: 2006-2039.

Sterne, T. E. 1954. Some remarks on confidence or fiducial limits. *Biometrika* 41: 275-278.

Stevens, S. S. 1951. Mathematics, measurement, and psychophysics. Pp. 1-49 in *Handbook of Experimental Psychology*, ed. S. S. Stevens. New York: Wiley.

Stevens, W. L. 1950. Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* 37: 117-129.

Stigler, S. 1986. *The History of Statistics: The Measurement of Uncertainty before* 1900. Cambridge, MA: Harvard University Press.

Stigler, S. 1994. Citation patterns in the journals of statistics and probability. *Statist. Sci.* 9: 94-108.

Stigler, S. 1999. *Statistics on the Table*. Cambridge, MA: Harvard University Press.

Stiratelli, R., N. Laird, and J. H. Ware. 1984. Random-effects models for serial observations with binary response. *Biometrics* 40: 1025-1035.

Stokes, M. E., C. S. Davis, and G. G. Koch. 2000. *Categorical Data Analysis Using the SAS System*, 2nd ed. Cary, NC: SAS Institute.

Strawderman, R. L., and M. T. Wells. 1998. Approximately exact inference for the common odds ratio in several 2 ×2 tables. *J. Amer. Statist. Assoc.* 93: 1294-1307.

Stuart, A. 1955. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* 42: 412-416.

Stukel, T. A. 1988. Generalized logistic models. *J. Amer. Statist. Assoc.* 83: 426-431.

Suissa, S., and J. J. Shuster. 1984. Are uniformly most powerful unbiased tests really best? *Amer. Statist.* 38: 204-206.

Suissa, S., and J. J. Shuster. 1985. Exact unconditional samples sizes for the 2 by 2 binomial trial. *J. Roy. Statist. Soc. Ser. A* 148: 317-327.

Suissa, S., and J. J. Shuster. 1991. The 2 ×2 matched-pairs trial: Exact unconditional design and analysis. *Biometrics* 47: 361-372.

Sundberg, R. 1975. Some results about decomposable (or Markov-type) models for multidimensional contingency tables: Distribution of marginals and partitioning of tests. *Scand. J. Statist.* 2: 71-79.

Tango, T. 1998. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statist. Medic.* 17: 891-908.

Tanner, M. A., and M. A. Young. 1985. Modelling agreement among raters. *J. Amer. Statist. Assoc.* 80: 175-180.

Tarone, R. E. 1985. On heterogeneity tests based on efficient scores. *Biometrika* 72: 91-95.

Tarone, R. E., and J. J. Gart. 1980. On the robustness of combined tests for trends in proportions. *J. Amer. Statist. Assoc.* 75: 110-116.

Tarone, R. E., J. J. Gart, and W. W. Hauck. 1983. On the asymptotic relative efficiency of certain noniterative estimators of a common relative risk or odds ratio. *Biometrika* 70: 519-522.

Tavaré, S., and P. M. E. Altham. 1983. Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics. *Biometrika* 70: 139-144.

Ten Have, T. R. 1996. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* 52: 473-491.

Ten Have, T. R., and A. R. Localio. 1999. Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics* 55: 1022-1029.

Ten Have, T. R., and A. Morabia. 1999. Mixed effects models with bivariate and univariate association parameters for longitudinal bivariate binary response data. *Biometrics* 55:85-93.

Ten Have, T. R., and D. H. Uttal. 1994. Subject-specific and population-averaged continuation ratio Logit models for multiple discrete time survival profiles. *Appl. Statist.* 43: 371-384.

Theil, H. 1969. A multinomial extension of the linear Logit model, *Internat. Econ. Rev.* 10: 251-259.

Theil, H. 1970. On the estimation of relationships involving qualitative variables. *Amer. J. Sociol.* 76: 103-154.

Thompson, R., and R. J. Baker. 1981. Composite link functions in generalized linear models. *Appl. Statist.* 30: 125-131.

Thompson, W. A. 1977. On the treatment of grouped observations in life studies. *Biometrics* 33: 463-470.

Thurstone, L. L. 1927. The method of paired comparisons for social values. *J. Abnormal Social Psych.* 21: 384-400.

Tjur, T. 1982. A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scand. J.*

Statist. 9: 23-30.

Tocher, K. D. 1950. Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika* 37: 130-144.

Toledano, A., and C. Gatsonis. 1996. Ordinal regression methodology for ROC curves derived from correlated data. *Statist. Medic.* 15: 1807-1826.

Train, K. 1986. *Qualitative Choice Analysis: Theory, Econometrics, and an Application.* Cambridge, MA: MIT Press.

Tsiatis, A. A. 1980. A note on the goodness-of-fit test for the logistic regression model. *Biometrika* 67: 250-251.

Tutz, G. 1989. Compound regression models for ordered categorical data. *Biometrical J.* 31: 259-272.

Tutz, G. 1991. Sequential models in categorical regression. *Comput. Statist. Data Anal.* 11: 275-295.

Tutz, G., and W. Hennevogl. 1996. Random effects in ordinal regression models. *Comput. Statist. Data Anal.* 22: 537-557.

Uebersax, J. S. 1993. Statistical modeling of expert ratings on medical treatment appropriateness. *J. Arner. Statist. Assoc.* 88: 421-427.

Uebersax, J. S., and W. M. Grove. 1990. Latent class analysis of diagnostic agreement. *Statist. Medic.* 9: 559-572.

Uebersax, J. S., and W. M. Grove. 1993. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics* 49: 823-835.

van der Heijden, P. G. M., and J. de Leeuw. 1985. Correspondence analysis: A complement to log-linear analysis. *Psychometrika* 50: 429-447.

van der Heijden, P. G. M., A. de Falguerolles, and J. de Leeuw. 1989. A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Appl. Statist.* 38: 249-292.

Verbeke, G., and E. Lesaffre. 1996. A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* 91: 217-221.

Verbeke, G., and G. Molenberghs. 2000. *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.

Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* 54: 426-482.

Walker, S. H., and D. B. Duncan. 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54: 167-179.

Walley, P. 1996. Inferences from multinomial data: Learning about a bag of marbles. *J. Roy. Statist. Soc. Ser. B* 58: 3-34.

Wardrop, R. L. 1995. Simpson's paradox and the hot hand in basketball. *Amer. Statist.* 49:24-28.

Ware, J. H., S. Lipsitz, and F. E. Speizer. 1988. Issues in the analysis of repeated categorical outcomes. *Statist. Medic.* 7: 95-107.

Watson, G. S. 1956. Missing and "mixed up" frequencies in contingency tables. *Biometrics* 12: 47-50.

Watson, G. S. 1959. Some recent results in chi-square goodness-of-fit tests. *Biometrics* 15: 440-468.

Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61: 439-447.

Wedderburn, R. W. M. 1976. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* 63: 27-32.

Wermuth, N. 1976. Model search among multiplicative models. *Biometrics* 32: 253-263.

Wermuth, N. 1987. Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable. *J. Roy. Statist. Soc. Ser. B* 49: 353-364.

Westfall, P. H., and R. D. Wolfinger. 1997. Multiple tests with discrete distributions. *Amer. Statist.* 51: 3-8.

Westfall, P. H., and S. S. Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* New York: Wiley.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1-26.

White, A. A., J. R. Landis, and M. M. Cooper. 1982. A note on the equivalence of several marginal homogeneity test criteria for categorical data. *Internat. Statist. Rev.* 50: 27-34.

Whitehead, J. 1993. Sample size calculations for ordered categorical data. *Statist. Medic.* 12: 2257-2271.

Whittaker, J. 1990. *Graphical Models in Applied Multivariate Statistics.* New York: Wiley.

Whittaker, J., and M. Aitkin. 1978. A flexible strategy for fitting complex log-linear models. *Biometrics* 34: 487-495.

Whittemore, A. S. 1978. Collapsibility of multidimensional tables. *J. Roy. Statist. Soc. Ser. B* 40: 328-340.

Whittemore, A. S. 1981. Sample size for logistic regression with small response probability. *J. Amer. Statist. Assoc.* 76: 27-32.

Wilks, S. S. 1935. The likelihood test of independence in contingency tables. *Ann. Math. Statist.* 6: 190-196.

Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* 9: 60-62.

Williams, D. A. 1975. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* 31: 949-952.

Williams, D. A. 1982. Extra-binomial variation in logistic linear models. *Appl. Statist.* 31: 144-148.

Williams, D. A. 1987. Generalized linear model diagnostics using the deviance and single-case deletions. *Appl. Statist.* 36: 181-191.

Williams, D. A. 1988. Comments on "The impact of litter effects on dose-response modeling in teratology." *Biometrics* 44: 305-308.

Williams, E. J. 1952. Use of scores for the analysis of association in contingency tables. *Biometrika* 39: 274-289.

Williams, O. D., and J. E. Grizzle. 1972. Analysis for contingency tables having ordered response categories. *J. Amer. Statist. Assoc.* 67: 55-63.

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.* 22: 209-212.

Wolfinger, R., and M. O'Connell. 1993. Generalized linear

mixed models: A pseudo-likelihood approach. *J. Statist. Comput. Simul.* 48: 233-243.

Wong, G. Y., and W. M. Mason. 1985. The hierarchical logistic regression model for multilevel analysis. *J. Amer. Statist. Assoc.* 80: 513-524.

Woolf, B. 1955. On estimating the relation between blood group and disease. *Ann. Human Genet.* (*London*) 19: 251-253.

Woolson, R. F., and W. R. Clarke. 1984. Analysis of categorical incomplete longitudinal data. *J. Roy. Statist. Soc. Ser. A* 147: 87-99.

Wu, C. F. J. 1985. Efficient sequential designs with binary data. *J. Amer. Statist. Soc.* 80: 974-984.

Yang, I., and M. P. Becker. 1997. Latent variable modeling of diagnostic accuracy. *Biometrics* 53: 948-958.

Yates, F. 1934. Contingency tables involving small numbers and the χ^2 test. *J. Roy. Statist. Soc. Suppl.* 1: 217-235.

Yates, F. 1948. The analysis of contingency tables with grouping based on quantitative characters. *Biometrika* 35: 176-181.

Yates, F. 1984. Tests of significance for 2 × 2 contingency tables. *J. Roy. Statist. Soc. Ser. A* 147: 426-463.

Yee, T. W., and C. J. Wild. 1996. Vector generalized additive models. *J. Roy. Statist. Soc. Ser. B* 58: 481-493.

Yerushalmy, J. 1947. Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. *Public Health Rep.* 62: 1432-1449.

Yule, G. U. 1900. On the association of attributes in statistics. *Philos. Trans. Roy. Soc. London Ser. A* 194:257-319.

Yule, G. U. 1903. Notes on the theory of association of attributes in statistics. *Biometrika* 2: 121-134.

Yule, G. U. 1906. On a property which holds good for all groupings of a normal distribution of frequency for two variables, with application to the study of contingency tables for the inheritance of unmeasured qualities. *Proc. Roy. Soc. Ser A* 77: 324-336.

Yule, G. U. 1912. On the methods of measuring association between two attributes. *J. Roy. Statist. Soc.* 75: 579-642.

Zeger, S. L., and M. R. Karim. 1991. Generalized linear models with random effects: A Gibbs sampling approach. *J. Amer. Statist. Assoc.* 86:79-86.

Zeger, S. L., K.-Y. Liang, and P. S. Albert. 1988. Models for longitudinal data: A generalized estimating equation approach. *Biometrics* 44: 1049-1060.

Zelen, M. 1971. The analysis of several 2 × 2 contingency tables. *Biometrika* 58: 129-137.

Zelen, M. 1991. Multinomial response models. *Comput. Statist. Data Anal.* 12: 249-254.

Zellner, A., and P. E. Rossi. 1984. Bayesian analysis of dichotomous quantal response models. *J. Economet.* 25: 365-393.

Zelterman. D. 1987. Goodness-of-fit tests for large sparse multinomial distributions. *J. Amer. Statist. Soc.* 82: 624-629.

Zermelo, E. 1929. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Math. Z.* 29: 436-460.

Zhang, H., J. Crowley, H. Sox, and R. Olshen. 1998. Tree-structured statistical methods. Pp. 4561-4573 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Zheng, B., and A. Agresti. 2000. Summarizing the predictive power of a generalized linear model. *Statist. Medic.* 19: 1771-1781.

Zhu, Y., and N. Reid. 1994. Information, ancillarity, and sufficiency in the presence of nuisance parameters. *Canad. J. Statist.* 22: 111-123.

例子索引*

* 索引部分(包括例子索引和主题索引)的页码为英文版页码。本书英文版各章页码范围如下:第 1 章 1—35 页,第 2 章 36—69 页,第 3 章 70—114 页,第 4 章 115—164 页,第 5 章 165—210 页,第 6 章 211—266 页,第 7 章 267—313 页,第 8 章 314—356 页,第 9 章 357—408 页,第 10 章 409—454 页,第 11 章 455—490 页,第 12 章 491—537 页,第 13 章 538—575 页,第 14 章 576—599 页,第 15 章 600—618 页,第 16 章 619—631 页。有需要查找索引的读者可以参照英文页码在本书相应处查阅。

主题索引

万卷方法总书目

万卷方法是我国第一套系统介绍社会科学研究方法的大型丛书，来自中国社科院、北京大学等研究机构和高校的两百余名学者参与了丛书的写作和翻译工作。至今已出版图书85个品种，其中绝大多数是2008年以来出版的新书。

85 社会科学方法论(国家十二五规划教材)
978-7-5624-6204-0

84 田野工作的艺术
978-7-5624-6257-6

83 图解 AMOS 在学术研究中的应用
978-7-5624-6223-1

82 应用 STATA 做统计分析(更新至 STATA10.0)
978-7-5624-4483-1

81 社会调查设计与数据分析——从立题到发表
978-7-5624-6074-9

80 质性研究导引
978-7-5624-6132-6

79 APA 格式——国际社会科学学术写作规范手册
978-7-5624-6105-0

78 如何做心理学实验
978-7-5624-6151-7

77 话语分析导论
978-7-5624-6075-6

76 心理学学位论文写作全程指导
978-7-5624-6113-5

75 心理学研究方法导论
978-7-5624-5828-9

74 分类数据分析
978-7-5624-6133-3

73 结构方程模型:AMOS 的操作与应用(附光盘版)
978-7-5624-5720-6

72 AMOS 与研究方法(第2版)
978-7-5624-5569-1

71 爱上统计学(第2版)
978-7-5624-5891-3

70 社会科学定量研究的变量类型、方法选择与范例解析
978-7-5624-5714-5

69 案例研究:设计与方法(中译第2版)
978-7-5624-5732-9

68 问卷设计手册:市场研究、民意调查、社会调查、健康调查指南
978-7-5624-5597-4

67 广义潜变量模型:多层次、纵贯性以及结构方程模型
978-7-5624-5393-2

66 调查问卷的设计与评估
978-7-5624-5153-2

65 心理学论文写作——基于 APA 格式的指导
978-7-5624-5354-3

64 心理学质性资料的分析
978-7-5624-5363-5

63 问卷统计分析实务:SPSS 操作与应用
978-7-5624-5088-7

62 如何做综述性研究
978-7-5624-5375-8

61 质性访谈方法
978-7-5624-5307-9

60 量表编制:理论与应用(校订新译本)
978-7-5624-5285-0

59 质性研究:反思与评论(第2卷)
978-7-5624-5143-3

58 实验设计原理:社会科学理论验证的一种路径
978-7-5624-5187-7

57 混合方法论:定性研究与定量研究的结合
978-7-5624-5110-5

56 社会统计学
978-7-5624-5253-9

55 校长办公室的那个人(质性研究个案阅读)
978-7-5624-4880-8

54 泰利的街角(质性研究个案阅读)
978-7-5624-4937-9

53 客厅即工厂(质性研究个案阅读)
978-7-5624-4886-0

52 标准化调查访问
978-7-5624-5062-7

51 解释互动论
978-7-5624-4936-2

50 如何撰写研究计划书
978-7-5624-5087-0

49 质性研究的理论视角:一种反身性的方法论
978-7-5624-4889-1

48 社会评估:过程、方法与技术
978-7-5624-4975-1

47 如何解读统计图表
978-7-5624-4906-5

46 公共管理定量分析:方法与技术(第2版)
978-7-5624-3640-9

45 量化研究与统计方法
978-7-5624-4821-1

44 心理学研究要义
978-7-5624-5098-6

43 调查研究方法(校订新译本)
978-7-5624-3289-0

42 分析社会情境:质性观察和分析方法
978-7-5624-4690-3

41 建构扎根理论:质性研究实践指南

978-7-5624-4747-4

40 参与观察法
978-7-5624-4616-3

39 文化研究:民族志方法与生活文化
978-7-5624-4698-9

38 质性研究方法:健康及相关专业研究指南
978-7-5624-4720-7

37 如何做质性研究
978-7-5624-4697-2

36 质性研究中的访谈:教育及社会科学研究者指南
978-7-5624-4679-8

35 案例研究方法的应用(中译第2版)
978-7-5624-3278-3

34 教育研究方法论探索
978-7-5624-4649-1

33 实用抽样方法
978-7-5624-4487-9

32 质性研究:反思与评论(第1卷)
978-7-5624-4462-6

31 社会科学研究的思维要素(第8版)
978-7-5624-4465-7

30 哲学史方法论十四讲
978-7-5624-4446-6

29 社会研究方法
978-7-5624-4456-5

28 质性资料的分析:方法与实践(第2版)
978-7-5624-4426-8

27 实用数据再分析法(第2版)
978-7-5624-4296-7

26 质性研究的伦理
978-7-5624-4304-9

25 叙事研究:阅读、倾听与理解
978-7-5624-4303-2

24 质化方法在教育研究中的应用(第2版)
978-7-5624-4349-0

23 复杂调查设计与分析的实用方法(第2版)
978-7-5624-4290-5

22 研究设计与写作指导:定性、定量与混合研究的路径
978-7-5624-3644-7

21 做自然主义研究:方法指南
978-7-5624-4259-2

20 多层次模型分析导论(第2版)
978-7-5624-4060-4

19 评估:方法与技术(第7版)
978-7-5624-3994-3

18 焦点团体:应用研究实践指南(第3版)
978-7-5624-3990-5

17 质的研究的设计:一种互动的取向(第2版)
978-7-5624-3971-4

16 组织诊断:方法、模型和过程(第3版)
978-7-5624-3055-1

15 民族志:步步深入(第2版)
978-7-5624-3996-7

14 分组比较的统计分析(第2版)
978-7-5624-3942-4

13 抽样调查设计导论(第2版)
978-7-5624-3943-1

12 定性研究(第3卷):经验资料收集与分析的方法(2版)
978-7-5624-3944-8

11 定性研究(第4卷):解释、评估与描述(第2版)
978-7-5624-3948-6

10 定性研究(第1卷):方法论基础(第2版)
978-7-5624-3851-9

9 定性研究(第2卷):策略与艺术(第2版)
978-7-5624-3286-9

8 社会网络分析法(第2版)
978-7-5624-2147-4

7 公共政策内容分析方法:
978-7-5624-3850-2

6 复杂性科学的方法论研究
978-7-5624-3825-0

5 社会科学研究:方法评论
978-7-5624-3689-8

4 论教育科学:基于文化哲学的批判与建构
978-7-5624-3641-6

3 科学决策方法:从社会科学研究到政策分析
7-5624-3669-0

2 电话调查方法:抽样、筛选与监控(第2版)
7-5624-3441-7

1 研究设计与社会测量导引(第6版)
978-7-5624-3295-1

为了建设好“万卷方法”，更好地服务学界，现由重庆大学出版社和人大经济论坛做出决定，凡购买重庆大学出版社的万卷方法系列图书的读者，填写以下信息调查表（复印即可），邮寄给我们（400030　重庆大学出版社　林佳木），经过认证后，我们将会赠送人大经济论坛币 100 个（可免费下载丛书相关学习资料并与教师及学友进行交流）：

读者情况调查表	
姓名	
单位	
联系电话	
E-mail	
论坛 ID	
使用书籍	
购买渠道	
对丛书建设的建议	
邮政地址（邮编）	

人大经济论坛

——国内最大的经济、管理、金融、统计类在线教育网站

人大经济论坛（网址：http://www.pinggu.org）依托中国人民大学经济学院，于 2003 年成立，致力于推动经济学科的进步，传播优秀教育资源，目前已经发展成为国内最大的经济、管理、金融、统计类的在线教育和咨询网站，也是国内最活跃和最具影响力的经济类网站。

1. 拥有国内经济类教育网站最多的关注人数，注册用户以百万计，日均数十万经济相关人士访问本站。

2. 是国内最丰富的经管类教育资源共享数据库和发布平台。

3. 论坛给所有会员提供学术交流与讨论的平台，同时也有网络社交 SNS 的空间，经管百科提供了丰富专业的经管类在线词典，数据定制和数据处理分析服务是您做实证研究的好帮手，免费的经济金融数据库使您不再为数据发愁，更有完善的经管统计类培训和教学相关软件，只要您是学习、研究或从事经管类行业，人大经济论坛就能满足您的需要！